




Doktorski studij
 "Farmaceutsko-biokemijske znanosti"
 Kolegij "Biostatistika"

Statistička obradba podataka: završna razmatranja

Mladen Petrovečki



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

Statistička hipoteza

- elementarna tvrdnja
- točna (istinita) ili netočna (neistinita)
- provjera hipoteze → **traženje istine**

⇒

Kako je Potjeh tražio istinu (www.bulaja.com)





Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

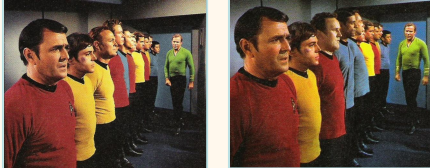
Statistička hipoteza

- istina – stvarno, objektivno stanje
- probabilistički sustav:
istina → **vjerojatnost**
- značajno – ono što se ostvaruje na svaki drugi način osim slučajno:
iskaz vjerojatnosti → **razina značajnosti**



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

Nulta hipoteza



Nema razlike!



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

Dokazivanje statističke hipoteze

- postavljanje hipoteze
- odabir statističkog testa
- određivanje razine značajnosti
- izračunavanje statistike testa
- zaključivanje



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

Pogrješke testiranja hipoteze

PRAVO STANJE		ZAKLJUČENO
RAZLIKA POSTOJI (H_1)	RAZLIKA NE POSTOJI (H_0)	
ISPRAVAN ZAKLJUČAK	α pogreška (I. vrste) (H_0 odbac.)	RAZLIKA POSTOJI
β pogreška (II. vrste)	ISPRAVAN ZAKLJUČAK (H_0 prihv.)	RAZLIKE NEMA



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

Uzorak i populacija



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Uzorak

- reprezentativan
- mjerljiv
- slučajni (probabilistički)

- jednostavni slučajni
- sustavni slučajni
- slojevit slučajni
- skupovni (klasterirani) slučajni



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Veličina uzorka

Small (<30) Sample Size Calculation

The formula for the sample size necessary to produce results accurate to a specified confidence and margin of error is:

$$n = \left(\frac{z_{\alpha/2} \sigma}{B} \right)^2$$

where:

$z_{\alpha/2}$ is known as the critical value, the positive z value that is at the vertical boundary for the area of $\alpha/2$ in the right tail of the standard normal distribution.

σ is the population standard deviation.

n is the sample size.

This formula can be used when you know σ and want to determine the sample size necessary to establish, with a confidence of $1 - \alpha$, the mean value μ to within $\pm B$.

As a general rule of thumb, if your sample size n is greater than 30, you can replace σ by the sample standard deviation s .

If your sample size is less than or equal to 30, the population must be normally distributed and you must know the population standard deviation σ in order to use the formula above. An alternate solution method is to use the Student t Distribution developed by William Gosset.

<http://www.isixsigma.com/>



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Određivanje veličine uzorka



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Analiza snage testa (power analysis)



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

<http://www.cs.uicw.edu/~rleth/Power/>



Randomizacija



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

<http://www.randomizer.org/>



Statistička i stvarna značajnost

- Statistička značajnost: ukazuje je li razlika ili povezanost između skupina značajna

Pitanje: je li važna?

- Odgovor: stvarna (klinička) značajnost i tumačenje znanstvenika



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Pogrješka temeljne prosudbe



Kahneman D, Tversky A.
On prediction and judgement.
Oregon Res Inst Bull 1972;371:100.



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Teorija... (1/3)

Postupak	Stanje	
	postoji (npr. bolestan)	ne postoji (npr. zdrav)
pozitivan nalaz	ISPRAVNO POZITIVNI (TP)	LAŽNO POZITIVNI (FP)
negativan nalaz	LAŽNO NEGATIVNI (FN)	ISPRAVNO NEGATIVNI (TN)



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Teorija... (2/3)

Osjetljivost testa = $TP / (TP + FN)$
Specifičnost testa = $TN / (FP + TN)$

Postupak	Stanje	
	postoji (npr. bolestan)	ne postoji (npr. zdrav)
pozitivan nalaz	ISPRAVNO POZITIVNI (TP)	LAŽNO POZITIVNI (FP)
negativan nalaz	LAŽNO NEGATIVNI (FN)	ISPRAVNO NEGATIVNI (TN)



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Teorija... (3/3)

Osjetljivost testa = $TP / (TP + FN)$

Specifičnost testa = $TN / (FP + TN)$

Pozitivna prediktivna vrijednost = $TP / (TP + FP)$

Negativna prediktivna vrijednost = $TN / (FN + TN)$

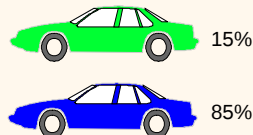
Postupak	Stanje	
	postoji (npr. bolestan)	ne postoji (npr. zdrav)
pozitivan nalaz	ISPRAVNO POZITIVNI (TP)	LAŽNO POZITIVNI (FP)
negativan nalaz	LAŽNO NEGATIVNI (FN)	ISPRAVNO NEGATIVNI (TN)



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Priča, saobraćajna nezgoda... (1/4)



PROMETNA NEZGODA, NOĆ, VOZAČ POBJEGNE
 $p(\text{zeleni}) = ?$

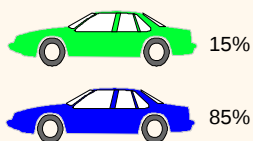
$p = 0.15$ (15%)



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Priča, saobraćajna nezgoda... (2/4)



OČEVIDAC NEZGODE: "ZELENI TAXI"
p(zeleni) = ?

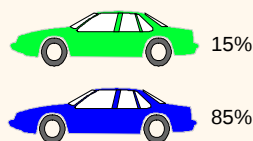
p = 1,0 (100%)



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Priča, saobraćajna nezgoda... (3/4)



OSPORAVANJE IZJAVE, TESTIRANJE NOĆNE VIDLJIVOSTI
Osjetljivost testa = 80%, Specifičnost testa = 80%

PREPOZNAJE 80%, GRIJEŠI 20% (OBJE)
p(zeleni) = ?

p = 0,8 (80%)



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Priča, saobraćajna nezgoda... (4/4)

Svjedok vidi	Stvarna boja vozila	
zeleno	12	17
plavo	03	68
ukupno	15	85

Osjetljivost testa = 80%, Specifičnost testa = 80%

p = 0,41 (41%)

Poz. PV = $12/(12+17) = 12/29 = 0,41$



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Korelacija, r

- 0 – matematička nepovezanost
- $\leq 0,25$ – nema povezanosti
- 0,26-0,50 – slaba povezanost
- 0,51-0,75 – umjerena povezanost
- $> 0,75$ – izvrsna povezanost
- 1 – matematička povezanost

Colton, 1974., *Statistics in Medicine*
(prema: Dawson & Trapp, 2001.)



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Pet "ne" r

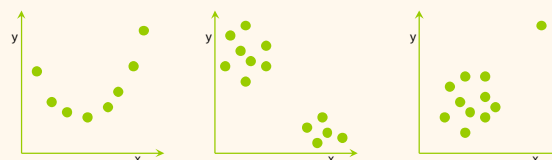
- r **nema** mjerne jedinice
- ne ovisi** o tome koji je pokazatelj x, a koji y
- $r=0$ znači: **nema** pravocrtna povezanosti
- ne smije** se rabiti za ekstrapolacije (samo interpolacije)
- ne označava** uzročno-posljedičnu vezu (samo povezanost)



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Kada NE računati r



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Članak, Biochem. Med.

- <http://www.ed/Vol17>
- by Ksenija

Što treba znati kada izračunavamo koeficijent korelacije?
What we need to know when calculating the coefficient of correlation?

Martina Ilić, Ksenija Babić, Ljiljana Zlatić, Mladen Petrović
*Katedra za medicinsku informatiku, Medicinski fakultet Sveučilišta u Rijeci, Rijeka, Croatia
*Zavod za laboratorijsku dijagnostiku, Klinički bolnički centar „Rijeka“, Rijeka
*Institute of Laboratory Diagnosis, Rijeka Clinical Hospital Center, Rijeka, Croatia
*Klinički zavod za laboratorijsku dijagnostiku, Klinički bolnički centar „Dubrava“, Zagreb
*Clinical Institute of Laboratory Diagnosis, Dubrava Clinical Hospital, Zagreb, Croatia

Sadržaj

Korelacija je statistički postupak za traživanje povezanosti dviju varijabli. Vrijednost korelacije brojeva se izraziti koeficijentom korelacije, najčešće Pearsonovim ili Spearmanovim, dok se značajnost korelacije izražava vrijednošću P. Koeficijent korelacije pokazuje u kojoj su mjeri promjene vrijednosti jedne varijable povezane s promjenama vrijednosti druge varijable. Predznak koeficijenta korelacije (+ ili -) govori nam o prirodi povezanosti. Povezanost označavanja korelacije vrijedno se pripisuje odnosu na uspjeh za traživanje korelacije. Koeficijent korelacije i značajnost korelacije, višestruke korelacije korelacije, poravnatost, utjecaj pojedinih varijabli, povezanost (koeficijent determinacije), te upotreba dva koeficijenta korelacije.

Ključne riječi: korelacija, Pearsonov koeficijent korelacije, Spearmanov koeficijent korelacije, koeficijent determinacije, pogreška, statistika

Abstract

Correlation is a statistical procedure applied to calculate association between two variables. The value of correlation is numerically shown by a coefficient of correlation, most often by Pearson's or Spearman's coefficient, while the significance of the coefficient is expressed by P value. The coefficient of correlation shows the extent to which changes in the value of one variable are correlated to changes in the value of the other. A sign preceding the coefficient of correlation (+ or -) indicates the direction of correlation. The most frequent error in calculating correlation are related to conditions for calculation, interpretation of the coefficient and correlation significance, high correlation coefficient, assumption of causal relationship, the strength of correlation coefficient of determination, and comparison of two correlation coefficients.

Key words: correlation, Pearson's correlation coefficient, Spearman's correlation coefficient, coefficient of determination, error, statistics



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

Ne zaboravite: 1-2-3-4!

1. Vrijednost P uvijek se piše s tri decimalna mjesta; nema "P<0,05", "P>0,05" ili p=NS
2. Vrijednost statističkog testa i koeficijenta korelacije pišu se s dva decimalna mjesta
3. Udjeli za N<100 i diskretne vrijednosti izražavaju se cijelim brojevima (bez decimalnih mjesta)
4. Učestalost u malim skupinama nije primjereno predstavljati relativno ("od tri miša 33% je pobjeglo")



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

Pogrješke...



http://www.cs.unc.edu/~pozefsky/COMP380_F06/FatalError.jpg



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

Članak, CMJ

- <http://www.cmj.hr/2004/4>
- by Tom Lang
(<http://www.tomlang.com>)

Error #5: Using the standard error of the mean (SEM) as a descriptive statistic or as a measure of precision for an estimate

The mean and standard deviation describe the center and variability of a normal distribution of a characteristic for a sample. The mean and standard error of the mean (SEM), however, are an estimate (the mean) and a measure of its precision (the SEM) for a characteristic of a population. However, the SEM is always smaller than the standard deviation, so it is sometimes reported instead of the standard deviation to make the measurements look more precise (16). Although the SEM is a measure of precision for an estimate (1 SEM on either side of the mean is essentially a 68% confidence interval), the preferred measure of precision in medicine is the 95% confidence interval (17). Thus, the mean and SEM can sometimes refer to a sample and sometimes to a population. To avoid confusion, the mean and standard deviation are the preferred summary statistics for (normally distributed) data, and the mean and 95% confidence interval are preferred for reporting an estimate and its measure of precision.

For example, if the mean weight of a sample of 30 subjects is 70.5 kg, the SEM is 1.2 kg, and the 95% confidence interval is 68.1 to 72.9 kg.



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

subjects was compared by grouping by sex, age, age at illness onset, and duration of illness. CFS subjects were stratified by sex, age (<40 years, 40–40 years, >50 years), onset type (gradual versus sudden) and duration of illness (<5 years, >5 years) to determine whether an association with autoantibodies existed.

Journal of Autoimmune Diseases 2005;2:1740



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

Table 1

Characteristics of patients in routine and inception cohorts

Characteristic	Routine cohort (cross-sectional)	Inception cohort (longitudinal)
Patients (n)	707	106
Age (years; mean ± SD)	54.1 ± 14.9	50.5 ± 15.6
Sex (% female)	79.9	75.2
Rheumatoid factor (% positive)	55.6	78.1
Disease duration at baseline (mean ± SD; range)	8.1 ± 10.6 years	11.5 ± 12.5 weeks
Duration of follow up (years; mean ± SD; range)	-	2.2 ± 1.3; 1–7.25
Disease activity characteristics (median [1st,3rd quartile])	At cross-section	At baseline
Swollen joint count (0–28)	3 (1,7)	7 (4,13)
Tender joint count (0–28)	2 (0,6)	8 (3,18)
ESR (mm; normal <20)	23 (14,55)	49 (24,70)
CRP (mg/dL; normal <1.0)	1.1 (0.5,2.7)	5.1 (1.5,17.0)
Patient assessment of pain (mm; 0–100)	37 (19,53)	50 (32,68)
Patient global assessment of activity (mm; 0–100)	37 (19,58)	51 (32,69)
Evaluator global assessment of activity (mm; 0–100)	34 (19,49)	44 (31,58)
HAQ (0–3)	0.875 (0.25;1.5)	0.75 (0.1,1.5)
Larsen score	-	1 (0,7)

Completeness of data for analysis

Cross-sectional correlation between composite indices (n [%])	767/767 (100)	105/106 (99.1%)
Cross-sectional correlation with HAQ scores (n [%])	720/767 (93.9)	104/106 (98.1%)
Discriminant validity, 1-year follow up (n [%])	-	91/106 (89.0%)
Construct validity, 3-year follow up (n [%])	-	56/80 (70.0%)

Completeness of data was the prerequisite for inclusion. *Used to validate the results from the cross-sectional analyses in the routine cohort. †Including complete radiological data. CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; HAQ, Health Assessment Questionnaire; SD, standard deviation.



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

Arthritis research & therapy 2005;7:R796

Table 1: Characteristics of subjects evaluated for autoantibodies.

	CFS	Non-Fatigued
Adjuvant Case-Control		
Subjects (n = 56)	22	34
Female (n = 52)	19	33
Male (n = 4)	3	1
Age Group (yrs)		
18-29 (n = 9)	3	6
30-39 (n = 18)	7	11
40-49 (n = 23)	12	11
50-59 (n = 6)	0	6
Mean Age	27 years	38 years
Mean Age Onset	36 years	46 years
Onset type		
Sudden	8	14
Gradual	14	20
Mean Illness Duration	69 months	128 months
Number of Subjects III		
< 5 years	10	14
> 5 years	12	23
Medica Foundation		
Subjects (n = 34)	37	57
Female (n = 64)	31	53
Male (n = 30)	6	4
Age Group (yrs)		
18-29 (n = 14)	1	13
30-39 (n = 18)	8	10
40-49 (n = 24)	13	13
50-59 (n = 36)	15	21
Mean Age	46 years	42 years
Mean Age Onset	36 years	42 years
Onset type		
Sudden	5	10
Gradual	12	21
Mean Illness Duration	128 months	128 months
Number of Subjects III		
< 5 years	14	14
> 5 years	23	23

of Autoimmune Diseases 2005;2:1740

were not shown to differ from those without the antibodies by other disease activity parameters including anti-dsDNA antibodies ($P = 0.99$), lymphocyte count ($P = 0.70$) and the global BILAG score ($P = 0.85$). However, patients with IgG anti-MBL antibodies had a higher ESR (38.7 mm/first hour) than those without (26.1 mm/first hour) ($P = 0.005$).

Patients with and without IgG anti-MBL antibodies were not significantly different in disease activity as defined by the BILAG score of other systems which

Lupus 2004;13:522

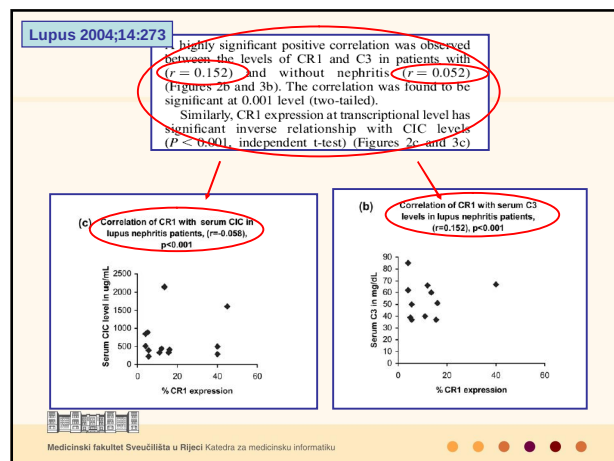
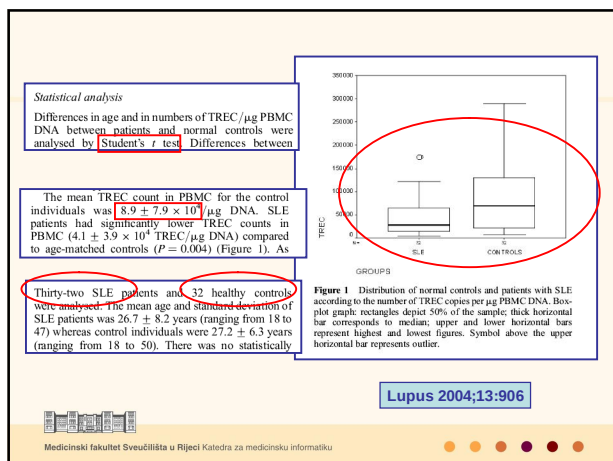


Table 1

SELDI-TOF MS protein peaks differentially expressed in paired sera from SLE before and after therapy

Mass (kDa)	P	Patients with visually distinct protein/total no. of patients	P	Patients with visually distinct protein/total no. of patients
4504	0.0001		0.005	
4708	0.0005		0.0006	
5739	0.0001		0.00000009	
6441	0.00008		0.0001	
6847	0.0005		0.003	
8510	0.0007	6/9	0.00001	12/15
9725	0.00004		0.00001	
11405	0.00001	8/9	0.00000002	14/15
11520	0.000007	8/9	0.000000001	14/15
11641	0.000005	8/9	0.000000001	14/15
11718	0.0004		0.00002	
11880	0.0002		0.0000006	
12703	0.0006	7/9	0.04	6/15
28816	0.0002	7/9	0.01	9/15
29389	0.0002		0.00005	
29827	0.0002		0.09	
36734	0.0002		0.00006	
32457	0.00001	6/9	0.00001	11/15
44837	0.0001	7/9	0.000009	11/15
74986	0.0002		0.00005	
76622	0.0004		0.0001	
76919	0.0008		0.00002	
79216	0.0007		0.00005	

Arthritis research & therapy 2005;7:R746

Nema ispitnih razdoblja!

Ispit

- kritička raščlamba statističkih postupaka objavljenog znanstvenog rada
- pismo (elektronički mladenp@medri.hr)
- poslati desetak dana prije usmene obrane
- časopis – citiran u *Current Contents-u*
- rad:
 - cjelokupno istraživanje (ne osvrta na drugi objavljeni rad!)
 - najmanje dva priloga (tablica, slika)
 - jasan opis statističkih postupaka

Ispit

- rad:
 - [primjer](#)
- kritička raščlamba statističkih postupaka objavljenog znanstvenog rada
 - [primjer](#)



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku



Top Ten Reasons to Become a Statistician

10. Deviation is considered normal.
9. We feel complete and sufficient.
8. We are mean lovers.
7. Statisticians do it discretely and continuously.
6. We are right 95% of the time.
5. We can safely comment on someone's posterior distribution.
4. We may not be normal but we are transformable.
3. We never have to say we are certain.
2. We are honestly significantly different.
1. No one wants our jobs.



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

<http://www.stetson.edu/~efriedma/mathumor.html>

