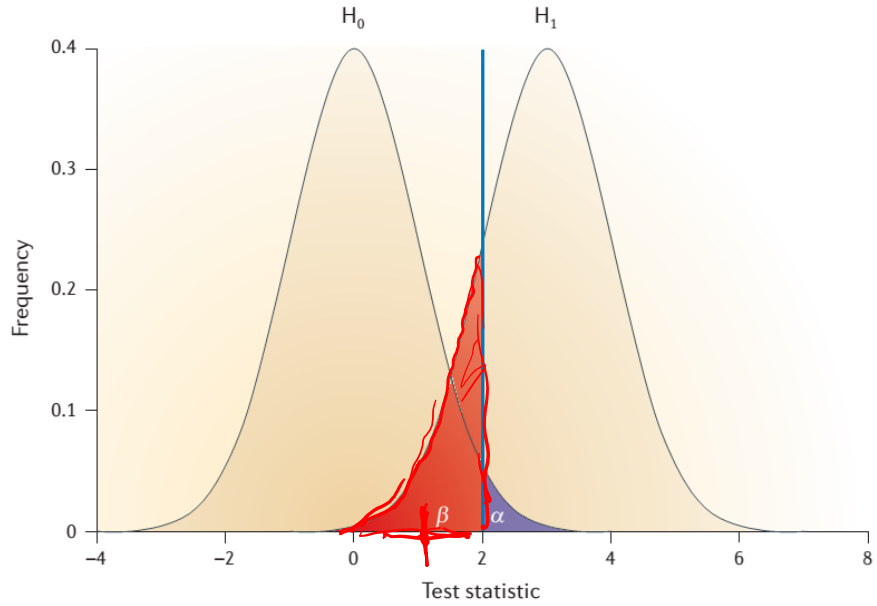
Testiranje statističkih hipoteza II

12.01.2024.

Pogreške u testiranju hipoteza

- **Pogreška tipa I.**
  - Odbacili smo nultu hipotezu koja je istinita
  - Vjerojatnost da ćemo počiniti pogrešku tipa I - α
  - Lažno pozitivni rezultati
- **Pogreška tipa II.**
  - Nismo odbacili nultu hipotezu koja je neistinita
  - Lažno negativni rezultati
  - Vjerojatnost da ćemo počiniti pogrešku tipa II - β
  - Vjerojatnost da nećemo počiniti pogrešku tipa II zove se **snaga testa**



Sham & Purcell, 2014, Nat Rev Genetics

	Nismo odbacili H_0	Odbacili smo H_0
H_0	-	Pogreška tipa I
H_1	Pogreška tipa II	-

Razina značajnosti – α

- Kada je rizik da ćemo napraviti grešku tipa I prevelik?
- Ovisi o okolnostima , uglavnom je prihvaćena vrijednost 5% (p = 0.05)
- Učestalost **greške tipa I** (broj greški tipa I na 100 eksperimenata) zove se *alfa razina*

Snaga testa

- Napravili smo t-test i zaključili da je razlika između srednje vrijednosti našeg uzorka i srednje vrijednosti populacije statistički značajna sa p-vrijednošću $p < 0.05$
- Ali što ako je $p > 0.05$?

- *Koji su mogući razlozi za $p > 0.05$?*
  - Nulta hipoteza je točna

  ili
  - Naš eksperiment nema dovoljno *statističke snage* i napravili smo **grešku tipa II**

Snaga testa

- Zašto p-vrijednost može biti veća od 0.05?
- Podsjetimo se:

$$t_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{SE_{\bar{x}}} \quad \text{i} \quad SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Snaga testa

- Zašto p-vrijednost može biti veća od 0.05?

1. Mala veličina učinka:

$\overline{X}$ je blizu srednje vrijednosti populacije

- Ne možemo razlikovati učinak tretmana od varijabilnosti u podacima
- Rješenje: Povećati veličinu učinka (npr. veća doza)

Snaga testa

- Zašto p-vrijednost može biti veća od 0.05?

2. „Šum” u podacima

s a zbog toga i $SE_{\bar{x}}$ je dosta velika

- Veliki broj u nazivniku će „poništiti” mali efekt
- Rješenje: Što je moguće više reducirati pogreške u mjerenjima

Snaga testa

- Zašto p-vrijednost može biti veća od 0.05?

3. Premala veličina uzorka

$SE_{\bar{x}}$ je dosta velika jer je $\sqrt{n}$ malen

- Veliki broj u nazivniku će „poništiti” mali efekt
- Rješenje: povećati broj uzoraka u eksperimentu

Snaga testa

- Koliko smo sigurni da smo mogli otkriti značajan učinak
- SNAGA $\propto \frac{veličina\ učinka\ \ i\ \alpha}{\sigma\sqrt{n}}$
- Točan izračun ovisi o vrsti statističkog testa i alternativnoj hipotezi

- Bitno: to što nismo odbacili nultu hipotezu ne znači da je nulta hipoteza točna!!!

Snaga testa

- Da bismo izračunali potrebnu veličinu uzorka moramo unaprijed odrediti željenu snagu (uglavnom 0.80 ili 0.90), razinu značajnosti α, veličinu učinka i procijeniti standardnu devijaciju
- Distribucija test statistike – ne-centralna t-distribucija, ovisi o ne-centralnom parametru ν i stupnjevima slobode

- Zašto ne uzmemo najveći mogući uzorak: troškovi, etički razlozi i biološka vs. statistička značajnost

Snaga testa

- Snaga testa ovisi o nekoliko različitih faktora
- Problem niske snage testa u biomedicinskim istraživanjima
- Prihvaćena vrijednost je obično 0.8

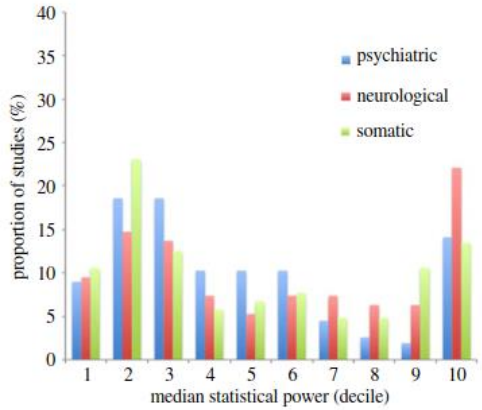
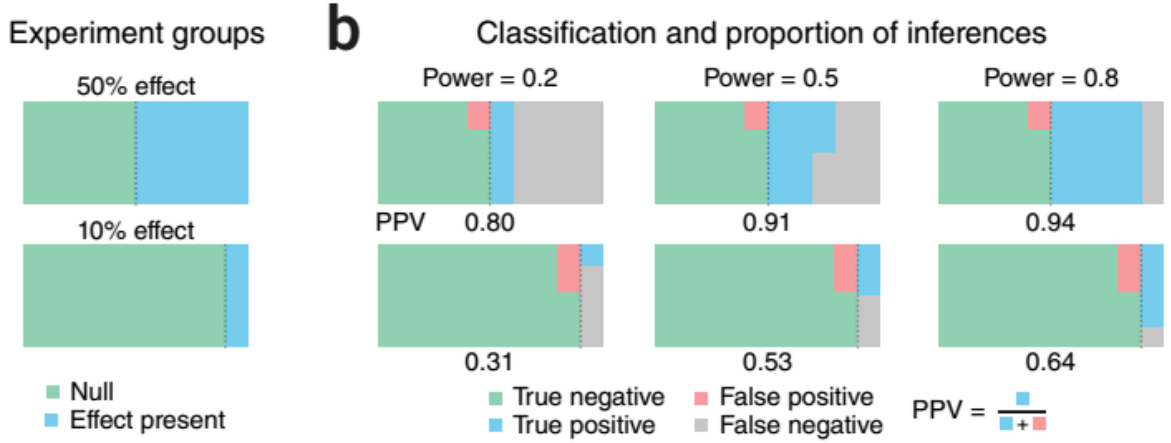


Figure 2. Distribution of statistical power of individual studies (sensitivity analysis). The distribution of the average statistical power of individual studies contributing to meta-analyses across three biomedical domains (psychiatry, neurology and somatic disease) is shown, restricted to meta-analyses indicating a statistically significant pooled effect size estimate only. This indicates a broadly similar bimodal distribution, albeit indicating higher average power overall. This overall pattern again appears to hold across all three domains of interest.

Dumas-Mallet E, R Soc Open Sci. 2017



Krzywinski & Altman, 2013, Nat Methods

Veličina učinka

- Koristi se kako bismo procijenili magnitudu učinka koji proučavamo

- Mogu se koristiti različite statistike:
  - Cohenov *d* (standardizirana razlika srednje vrijednosti)
  - Koeficijent korelacije (veza između kontinuiranih varijabli)
  - eta-kvadrat (ANOVA)
  - …..

Veličina učinka – razlika srednje vrijednosti

- Cohenov *d*

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}$$

- 0.2 – mali učinak, 0.5 – srednji učinak, 0.8 – veliki učinak

Alternative:

- Glassov *Δ* – uzorci s različitim varijancama (koristi samo varijancu kontrolne skupine)
- Hedgesov *g* – različite veličine uzoraka

Istraživanje treba biti reproducibilno i statistički ispravno

- 17–25% značajnih rezultata u društvenim znanostima (α = 0.05) je vjerojatno krivo, (Johnson, V. E. Proc. Natl Acad. Sci. USA, 2013) – nereproducibilnost rezultata je ozbiljan problem u znanosti
- Nature checklist:

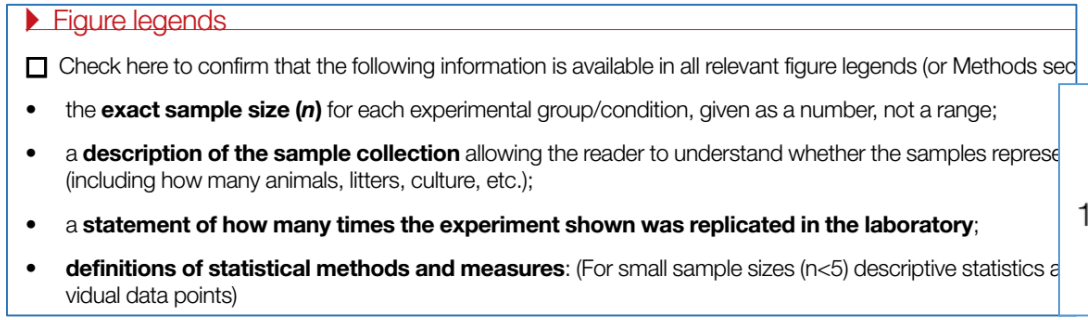
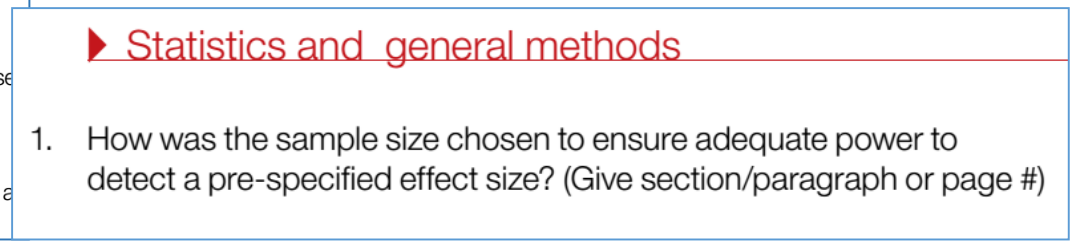


Figure legends

☐ Check here to confirm that the following information is available in all relevant figure legends (or Methods sec

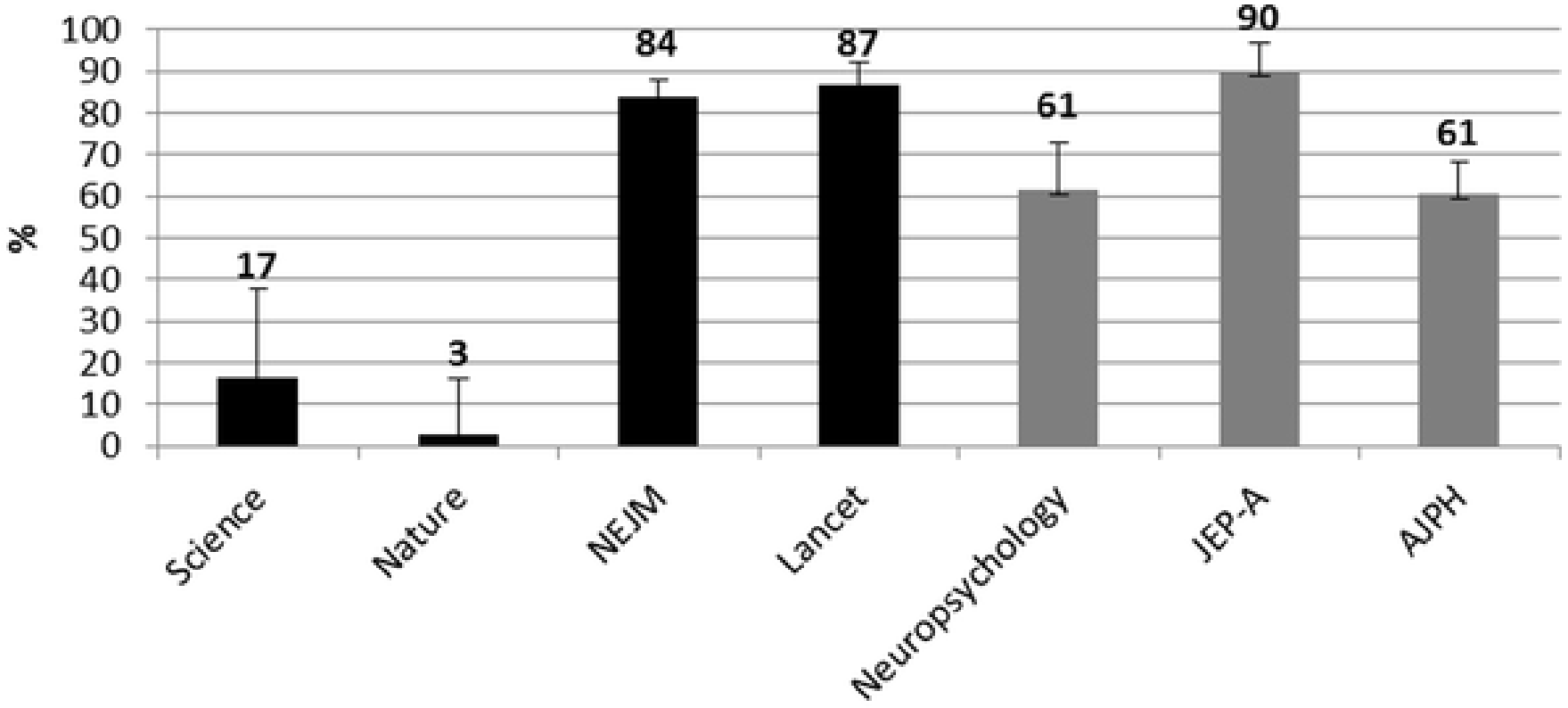
- the **exact sample size (*n*)** for each experimental group/condition, given as a number, not a range;
- a **description of the sample collection** allowing the reader to understand whether the samples represe (including how many animals, litters, culture, etc.);
- a **statement of how many times the experiment shown was replicated in the laboratory**;
- **definitions of statistical methods and measures**: (For small sample sizes (n<5) descriptive statistics a vidual data points)



Statistics and general methods

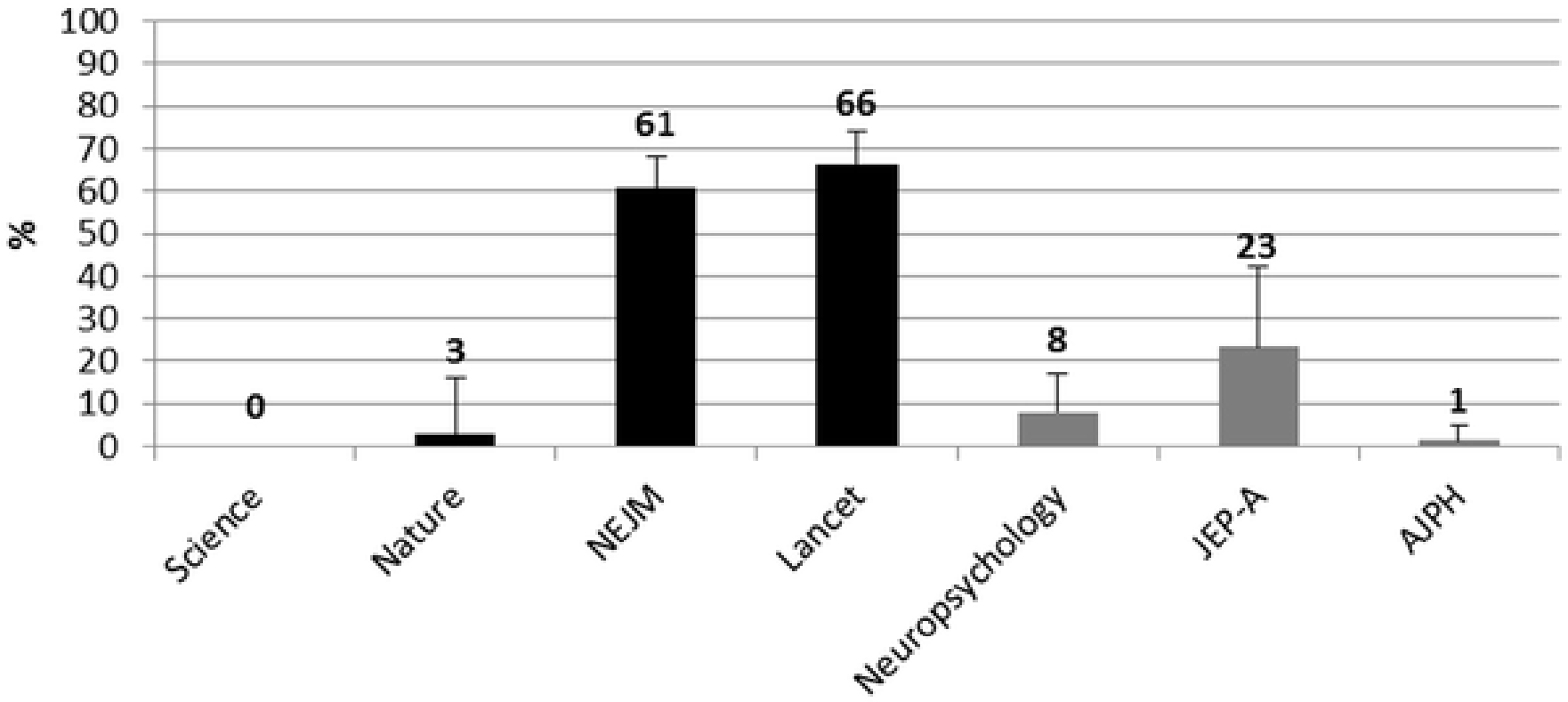
1. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size? (Give section/paragraph or page #)

Figure 2. Percentages of selected articles in each journal that reported a measure stated to be an effect size.



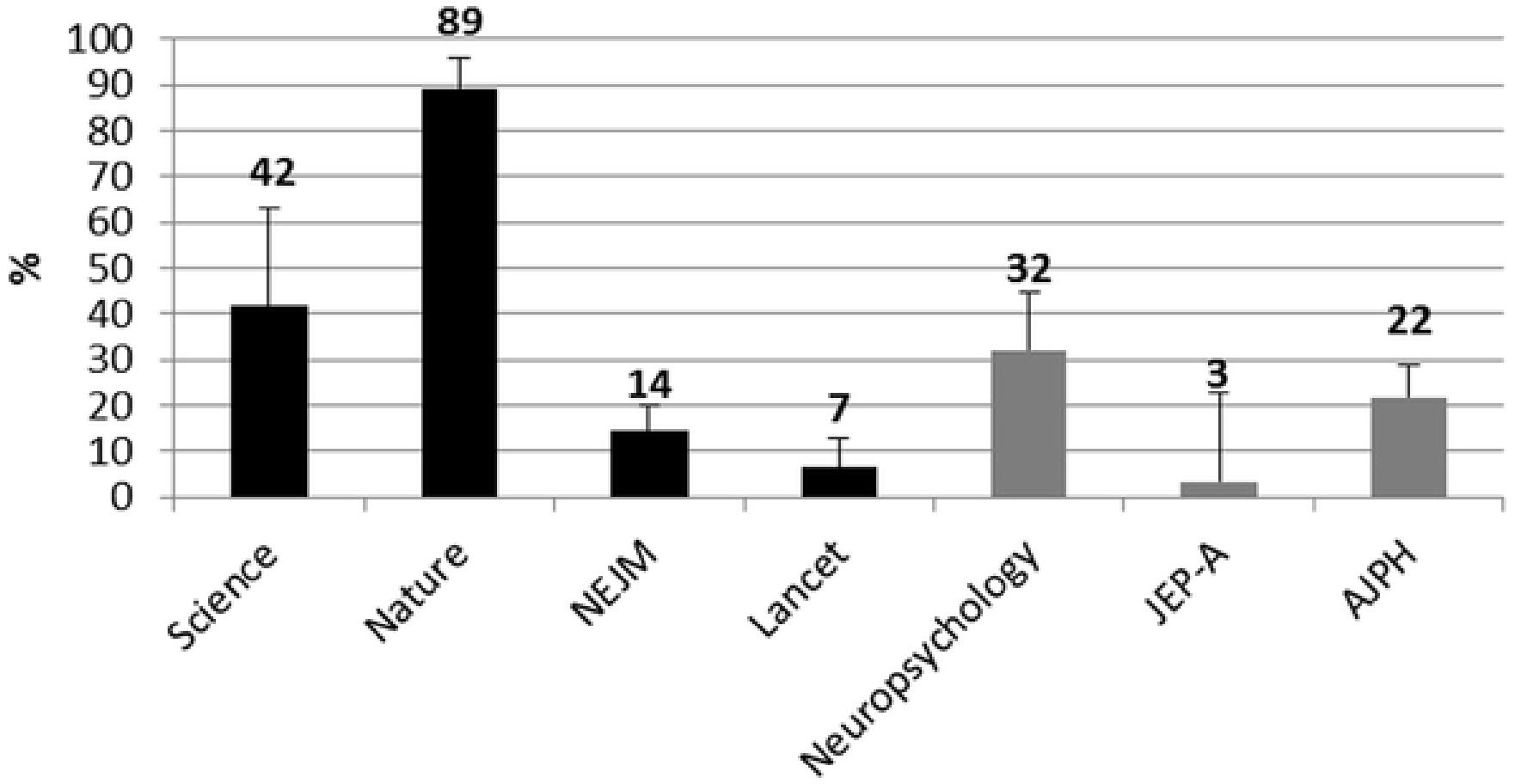
Tressoldi PE, Giofré D, Sella F, Cumming G (2013) High Impact = High Statistical Standards? Not Necessarily So. PLOS ONE 8(2): e56180. https://doi.org/10.1371/journal.pone.0056180
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0056180

Figure 5. Percentages of selected articles in each journal reporting a value of prospective power.



Tressoldi PE, Giofré D, Sella F, Cumming G (2013) High Impact = High Statistical Standards? Not Necessarily So. PLOS ONE 8(2): e56180. https://doi.org/10.1371/journal.pone.0056180
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0056180

Figure 6. Percentages of selected articles in each journal that used NHST without CI, ES or Model and Power estimation.



Tressoldi PE, Giofré D, Sella F, Cumming G (2013) High Impact = High Statistical Standards? Not Necessarily So. PLOS ONE 8(2): e56180. https://doi.org/10.1371/journal.pone.0056180
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0056180

Istraživanje treba biti reproducibilno i statistički ispravno

- Nature checklist:

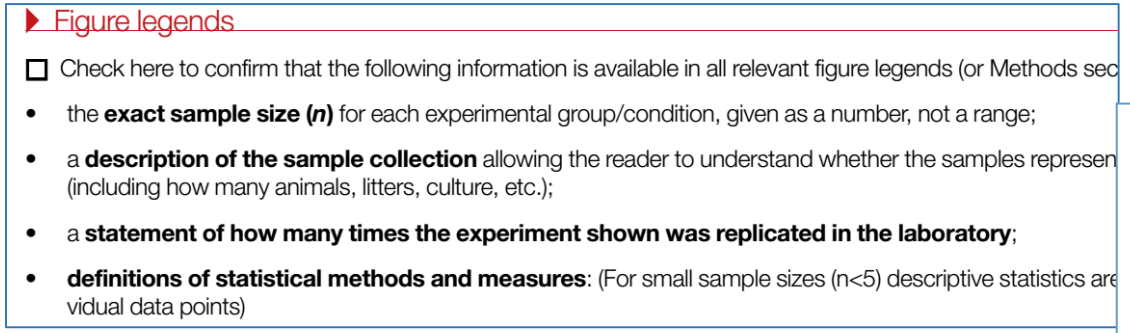
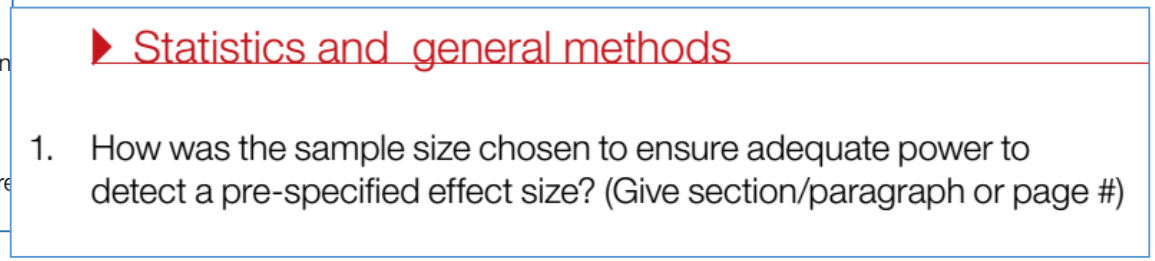


Figure legends

☐ Check here to confirm that the following information is available in all relevant figure legends (or Methods sec

- the **exact sample size (*n*)** for each experimental group/condition, given as a number, not a range;
- a **description of the sample collection** allowing the reader to understand whether the samples represen (including how many animals, litters, culture, etc.);
- a **statement of how many times the experiment shown was replicated in the laboratory**;
- **definitions of statistical methods and measures**: (For small sample sizes (n<5) descriptive statistics are vidual data points)



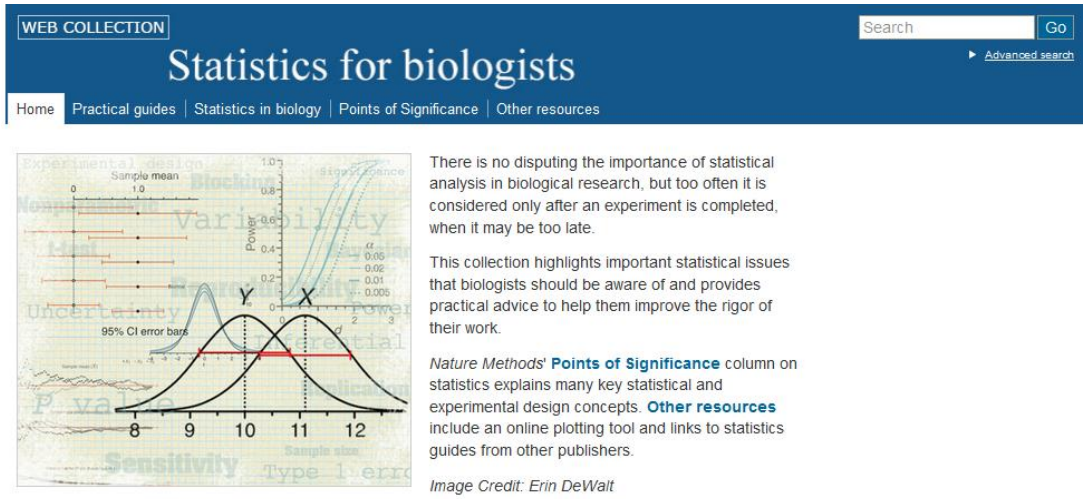
Statistics and general methods

1. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size? (Give section/paragraph or page #)

- Dizajnirate istraživanje tako da smanjite pogreške tipa I i tipa II (eksperimentalni dizajn)
- Odaberite prikladan statistički test (provjerite pretpostavke svog testa)

Nature statistics collection

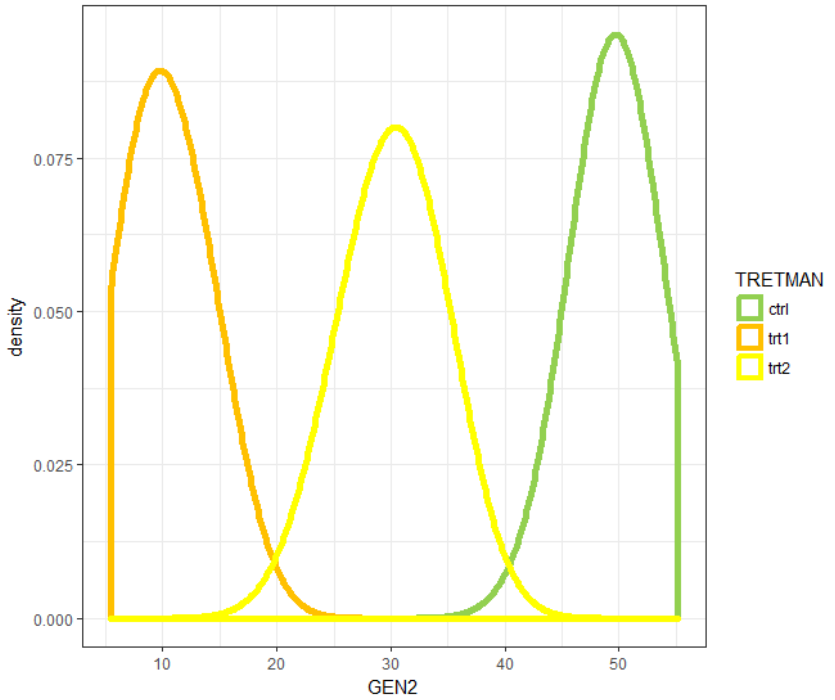
http://www.nature.com/collections/qghhqm



ANOVA (Analiza varijance)

- Istovremeno želimo testirati više od dvije skupine ispitanika

ID	STAROST	SPOL	TRETMAN	GEN1	GEN2	GEN3	STATUS
87	53	F	trt2	17.41	28.23	4.17	zdrav
119	53	M	ctrl	19.84	52.56	47.24	zdrav
67	52	M	trt2	19.19	27.49	12.07	zdrav
62	54	M	trt2	22.77	28.45	9.62	bolestan
131	55	F	ctrl	24.17	49.91	49.55	bolestan
50	54	F	trt1	17.15	15.32	10.67	zdrav
106	54	M	ctrl	17.92	44.95	51.39	zdrav
127	58	F	ctrl	20.06	53.19	49.71	bolestan
30	54	M	trt1	19.97	16.18	13.78	zdrav
72	54	F	trt2	25.44	27.58	11.81	zdrav



ANOVA (Analiza varijance)

- F-statistika – testira omjer varijabilnosti između skupina i unutar skupina

```
             Df Sum Sq Mean Sq F value Pr(>F)
TRETMAN       2  38784   19392    2207 <2e-16 ***
Residuals   147   1292       9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Post-hoc testovi – ako želimo odrediti između kojih skupina postoji statistički značajna razlika
- **Tukey’s HSD test**

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = GEN2 ~ TRETMAN, data = data)

$TRETMAN
                 diff       lwr       upr p adj
trt1-ctrl -39.38705 -40.79070 -37.98339     0
trt2-ctrl -19.52325 -20.92691 -18.11960     0
trt2-trt1  19.86379  18.46014  21.26745     0
```

Pretpostavke za korištenje ANOVE

- Varijabla koju testiramo je normalno distribuirana
- Uzorci imaju jednake varijance
- Uzorci su međusobno neovisni

- Neparametarske alternative:
- **Kruskal-Wallis test** – više od dva neovisna uzorka
- **Friedmanov test** – više od dva ovisna uzorka

Varijante ANOVE

- Proučavanje efekta dvije nezavisne kategoričke varijable
  - Interakcije među varijablama

- Analiza varijance ponovljenih uzoraka
  - Manji troškovi
  - Može dovesti do veće snage testa, jer varijabilnost unutar ispitanika može biti sistematska

Two-way ANOVA - primjer

- 150 ispitanika: 50 ctrl, 50 trt1, 50 trt2
- 50% ispitanika na posebnom režimu prehrane
- Zavisna varijabla: Ekspresija gena GEN2
- Postoji li utjecaj tretmana? Postoji li utjecaj prehrane?
- Two-way dizajn

ID	STAROST	SPOL	TRETMAN	GEN1	GEN2	GEN3	STATUS	DIET
144	55	F	ctrl	18.17	58.79	50.80	bolestan	NO
111	54	M	ctrl	20.56	48.08	49.61	bolestan	YES
113	56	F	ctrl	19.66	45.08	51.40	bolestan	NO
11	53	F	trt1	18.69	-0.96	6.00	zdrav	NO
91	56	F	trt2	19.41	30.05	6.03	bolestan	NO
64	56	M	trt2	18.95	11.28	11.71	zdrav	YES
63	50	F	trt2	21.84	36.94	11.26	zdrav	NO
149	50	M	ctrl	17.85	55.66	50.29	zdrav	NO
143	53	M	ctrl	18.71	48.40	45.89	bolestan	YES
38	55	M	trt1	23.62	9.71	14.97	bolestan	YES

Testiranje hipoteza

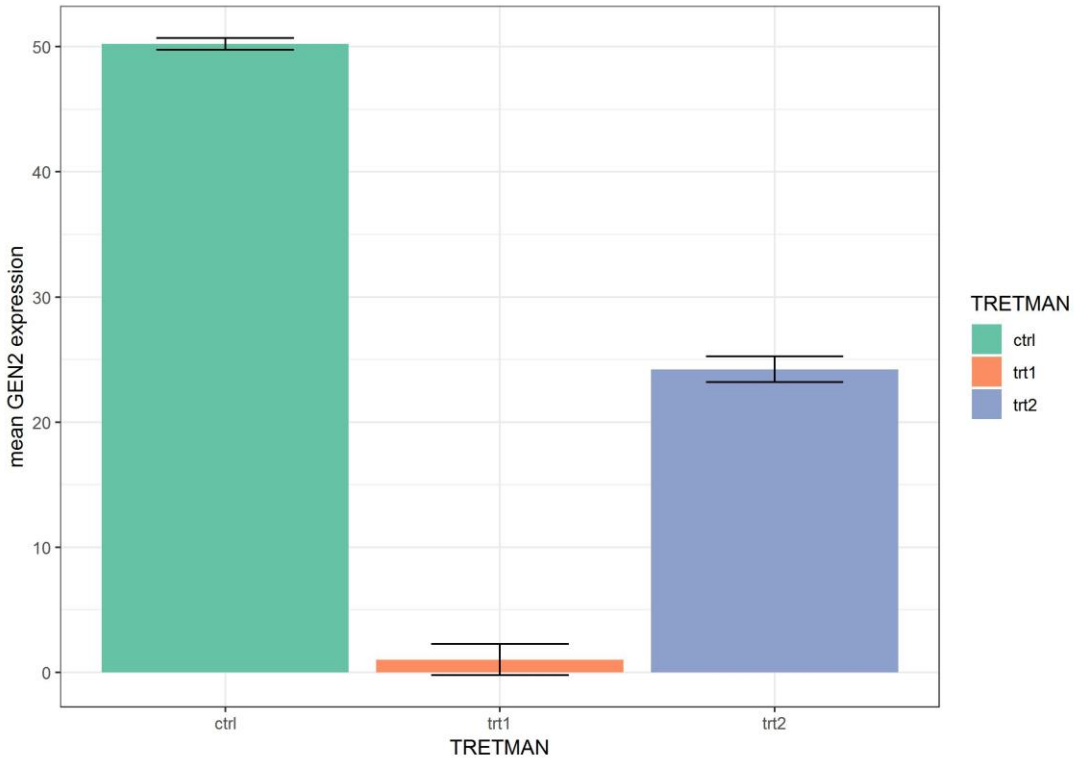
- Simultano testiramo 3 hipoteze:
  - Postoji li učinak tretmana na ekspresiju gena GEN2?
  - Postoji li učinak prehrane na ekspresiju gena GEN2?
  - Ovisi li učinak tretmana na ekspresiju gena GEN2 o prehrani (postoji li interakcija?)

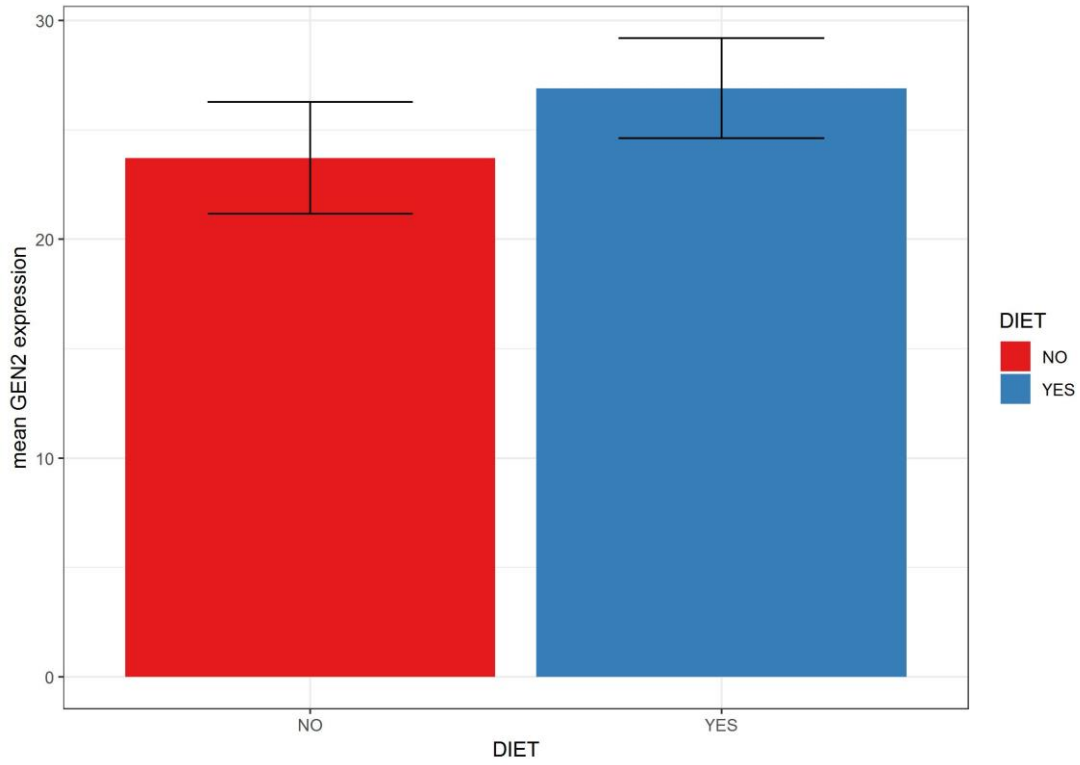
- 3 različita učinka
  - Glavni učinak
  - Interakcijski učinak
  - Jednostavni učinak

Naknadni testovi

- Naknadni testovi za značajne glavne učinke (post-hoc testovi)
- Naknadni testovi za značajne interakcijske učinke – analiza jednostavnih učinaka (učinak jedne varijable na svakoj razini druge varijable)

Glavni učinci

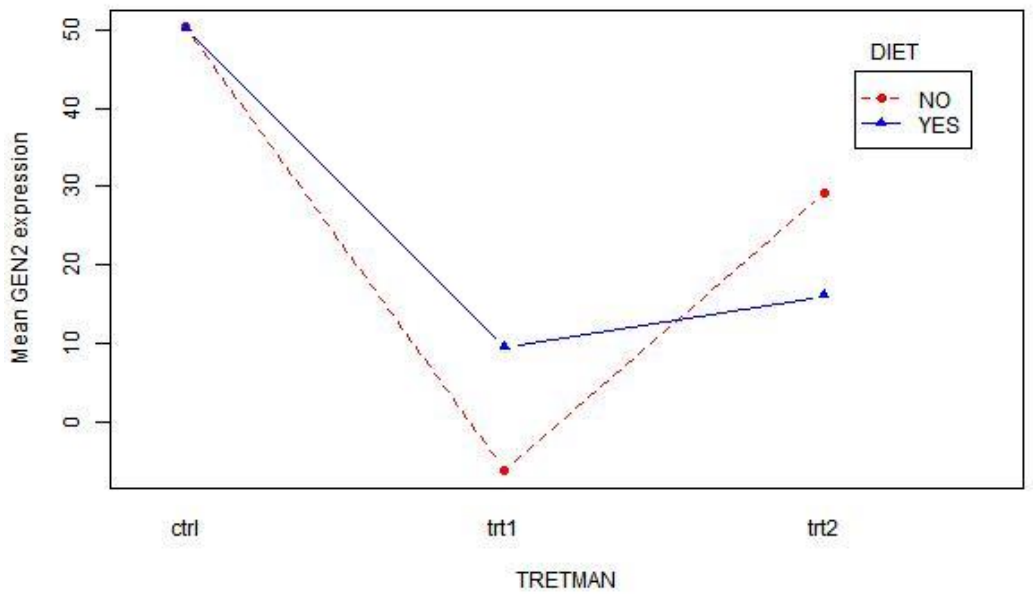




Prikazi interakcija

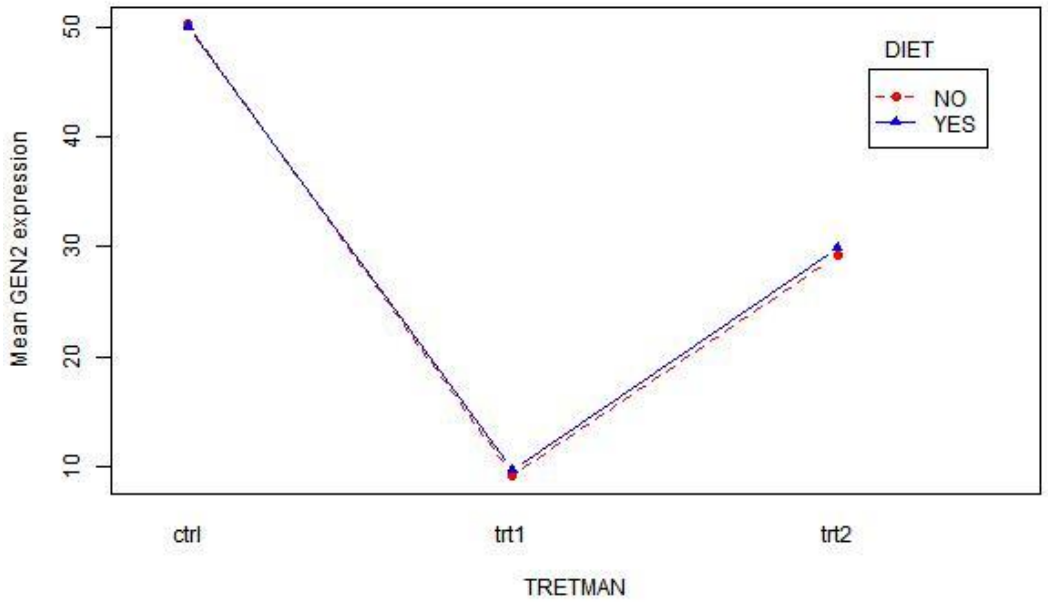
Naš primjer

```
> my.aov = aov(GEN2 ~ TRETMAN + DIET + TRETMAN:DIET,
+                     data = data)
> summary(my.aov)
              Df Sum Sq Mean Sq  F value Pr(>F)
TRETMAN        2  60620   30310 2393.450 <2e-16 ***
DIET           1     43      43    3.374 0.0683 .
TRETMAN:DIET   2   5110    2555  201.754 <2e-16 ***
Residuals    144   1824      13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Bez interakcija

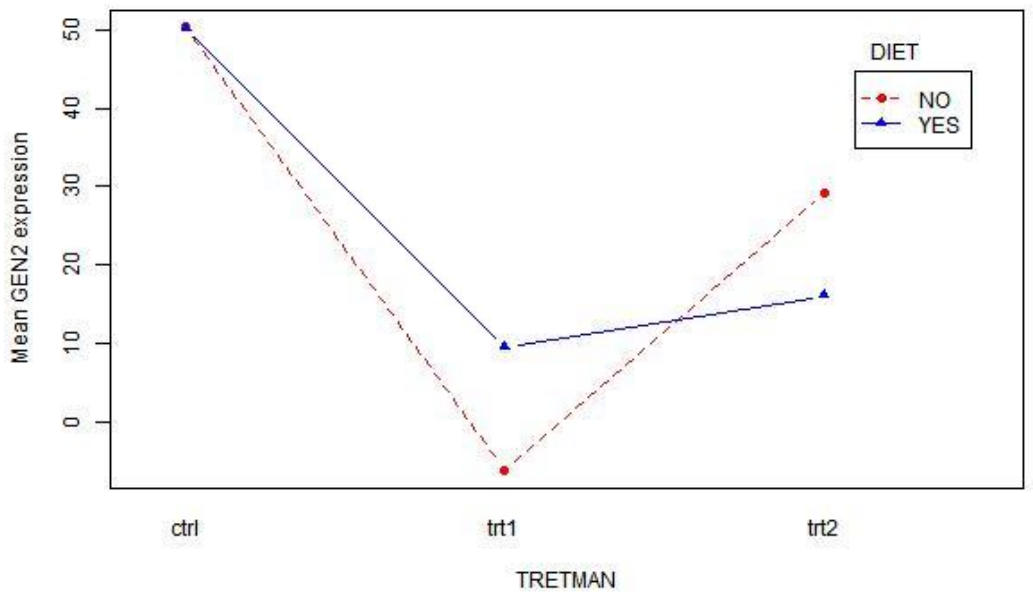
```
> my.aov = aov(GEN2 ~ TRETMAN + DIET + TRETMAN:DIET,
+                     data = data2)
> summary(my.aov)
              Df Sum Sq Mean Sq  F value Pr(>F)
TRETMAN        2  41898   20949 2032.427 <2e-16 ***
DIET           1      4       4    0.406  0.525
TRETMAN:DIET   2      6       3    0.269  0.765
Residuals    144   1484      10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Prikazi interakcija

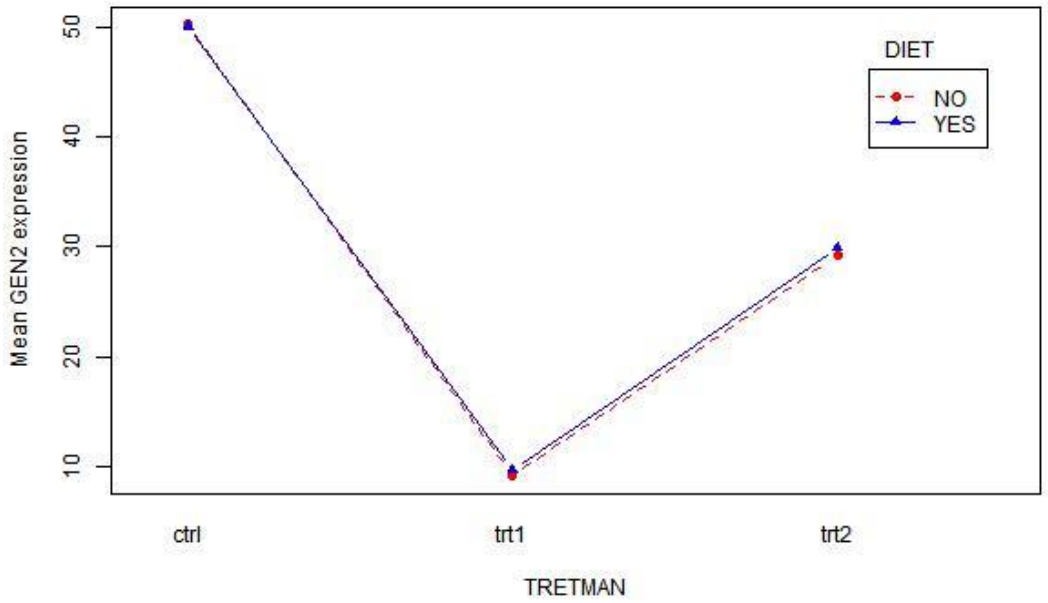
Naš primjer

```
> my.aov = aov(GEN2 ~ TRETMAN + DIET + TRETMAN:DIET,
+                     data = data)
> summary(my.aov)
              Df Sum Sq Mean Sq  F value Pr(>F)
TRETMAN        2  60620   30310 2393.450 <2e-16 ***
DIET           1     43      43    3.374 0.0683 .
TRETMAN:DIET   2   5110    2555  201.754 <2e-16 ***
Residuals    144   1824      13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Bez interakcija

```
> my.aov = aov(GEN2 ~ TRETMAN + DIET + TRETMAN:DIET,
+                     data = data2)
> summary(my.aov)
              Df Sum Sq Mean Sq  F value Pr(>F)
TRETMAN        2  41898   20949 2032.427 <2e-16 ***
DIET           1      4       4    0.406  0.525
TRETMAN:DIET   2      6       3    0.269  0.765
Residuals    144   1484      10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Hi-kvadrat test

- Postoji li veza između dvije kategoričke varijable?

		Status		
		Bolestan	Zdrav	
Primio tretman	NE	32	18	50
	DA	23	77	100
		55	95	150

- Usporedba očekivane i opažene frekvencije
- Izračun očekivanih frekvencija
- Izračun prikladne vrijednosti stupnjeva slobode

Hi-kvadrat test

- Kako izračunati očekivane frekvencije?

		Ishodi za varijablu STATUS		
		Bolestan	Zdrav	
Ishodi za varijablu TRETMAN	NE	$\frac{R1 \times C1}{TOTAL}$	$\frac{R1 \times C2}{TOTAL}$	R1
	DA	$\frac{R2 \times C1}{TOTAL}$	$\frac{R2 \times C2}{TOTAL}$	R2
		C1	C2	TOTAL

- Kako izračunati stupnjeve slobode?

$$df=(r-1)x(c-1)$$

r – broj redova

c – broj stupaca

Hi-kvadrat test

		Status		
		Bolestan	Zdrav	
Primio tretman	NE	32 (E = 18.33)	18 (E = 31.67)	50
	DA	23 (E = 36.67)	77 (E = 63.33)	100
		55	95	150

- Hi-kvadrat statistika = 22.396
- Stupnjevi slobode = 1
- P-vrijednost = 2.218×10^{-6}

Više od dva stupca i dva reda:

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

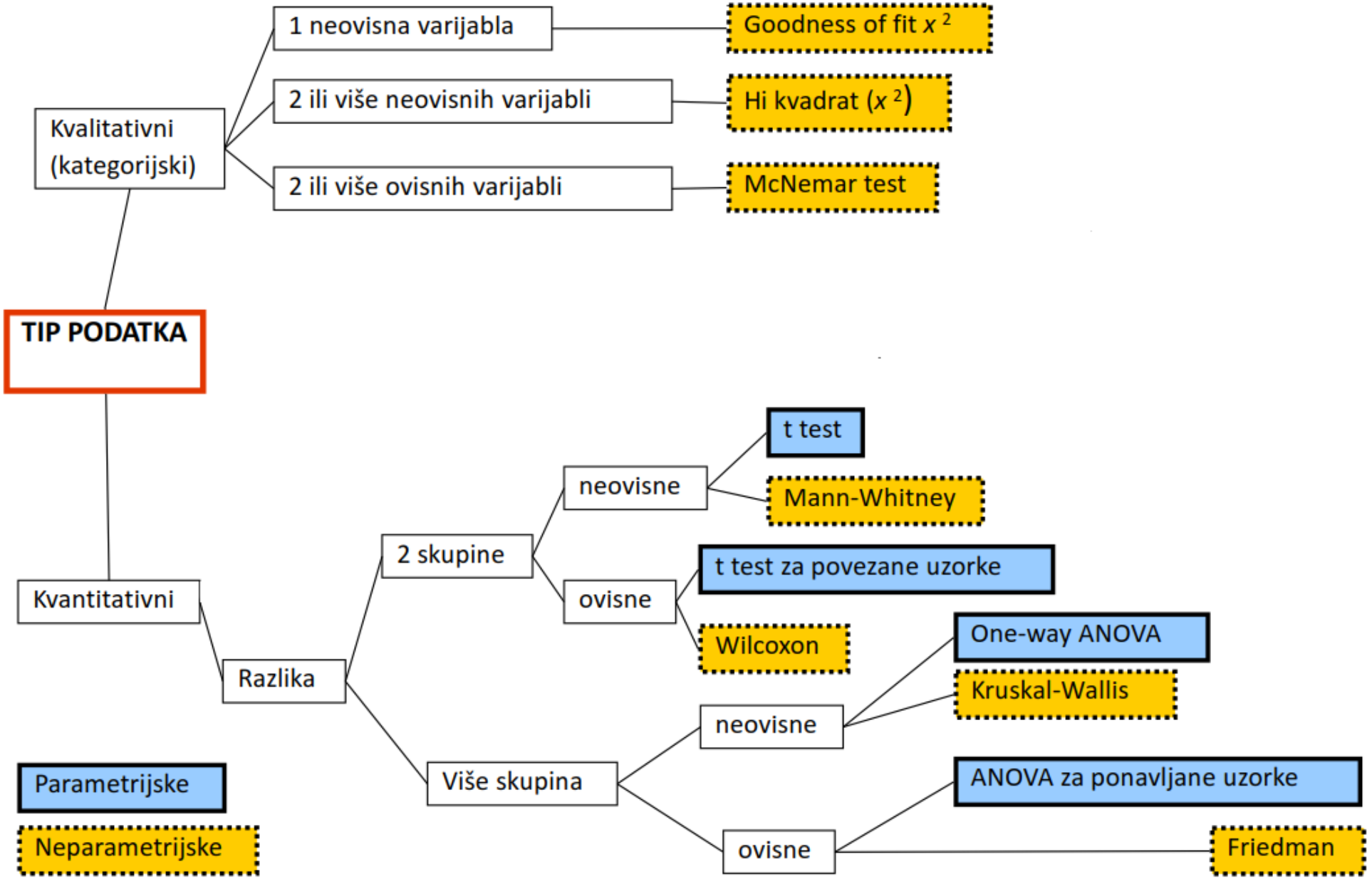
Dva stupca i dva reda:

$$\chi^2 = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Pretpostavke za korištenje Hi-kvadrat testa

- Kategoričke varijable su neovisne jedna o drugoj
- Očekivane vrijednosti E su dovoljno velike (>= 5)

- U slučaju međusobno ovisnih kategoričkih varijabli koristi se **McNemarov test**
- U slučaju malog broja uzoraka (E < 5) koristi se **Fisherov test**



Prilagođeno sa http://neuron.mefst.hr/docs/katedre/istrazivanja_bz/ibz%20dm1/DM1%20dan%203%20predavanje.pdf