Statistički testovi za kategoričke podatke

10.12.2024.

Rosa Karlić

Analize bioloških podataka 2024/2025

Primjer studije:

Kako izloženost zagađivaču iz vode (poliklorirani bifenili, PCB) utječe na imunološki odgovor vodozemaca i koji su daljnji učinci na dinamiku populacije i zdravlje ekosustava?

Ciljevi studije:

1. Istražiti promjene ekspresije gena u genima povezanim s imunološkim sustavom kod vodozemaca izloženih zagađenju (npr. CYP1A1, IL-6, IL-1β and TNF-α).
2. Istražiti imunološku reakciju (npr. razinu citokina) kod izloženih vodozemaca.
3. Kvantificirati razinu onečišćujućih tvari u vodi i procijeniti njihov ekološki utjecaj na ekosustav.
4. Mjeriti preživljenje, reprodukciju i ponašanje vodozemaca u zagađenim naspram kontrolnih okoliša.

Tipovi varijabli

Numeričke varijable

Diskretne ili kontinuirane

Kategoričke varijable

Nominalne ili ordinalne

Intervalne

Omjerne (*ratio*)

Intervalna varijabla je ona kod koje postoji poredak i razlika između dvije vrijednosti je smislena
Omjer dviju vrijednosti nema smislenu interpretaciju
Npr. Temperatura (F,°C),

Omjerna varijabla ima sva svojstva intervalne varijable i jasnu definiciju 0,0. Kada je varijabla jednaka 0,0, nema te varijable.
Npr. Aktivnost enzima, koncentracija tvari

Primjeri kategoričkih varijabli

Nominalne:

- Vrste stanica: matične stanice, diferencirane stanice, stanice raka.
- Vrste mutacija: tihe (silent), pogrešne (missense), besmislene (nonsense), pomak okvira čitanja (frameshift).
- Fenotipovi stanica: CD4+ T stanice, CD8+ T stanice, B stanice.
- Kategorije patogena: virusni, bakterijski, gljivični, parazitski.
- Tipovi staništa: Šuma, travnjak, močvara, pustinja.
- Način ishrane: biljojedi, mesojedi, svejedi.
- Izvori onečišćenja: industrijski, poljoprivredni, urbani.
- Klimatski pojasevi: tropski, umjereni, sušni, polarni.

Ordinalne:

- Razine ekspresije gena: Niska, srednja, visoka.
- Stadiji staničnog ciklusa: G1, S, G2, M.
- Ozbiljnost bolesti: blaga, umjerena, teška.
- Razine imunološkog odgovora: slab, umjeren, jak.
- Razine degradacije staništa: netaknuto, umjereno poremećeno, jako poremećeno.
- Kategorije gustoće naseljenosti: rijetka, umjerena, gusta.
- Kategorije indeksa kvalitete zraka: dobro, umjereno, nezdravo, opasno.
- Jačina onečišćenja: Niska, srednja, visoka.

Naš primjer studije:

Amphibian Responses to PCB Pollution						
Gene Expression	**Mutation Type**	**Immune Response**	**Activity Level**	**Survival Rate**	**Habitat Type**	**Pollution Severity**
Moderate	None	Adaptive	Decreased	Low	Pristine	Moderate
Low	Frameshift	Adaptive	Decreased	Low	Semi-polluted	Severe
Low	Missense	Adaptive	Increased	Medium	Polluted	Moderate
High	None	Innate	Decreased	High	Pristine	Moderate
High	Frameshift	Adaptive	Increased	Low	Polluted	Mild
Moderate	Missense	Adaptive	Increased	Medium	Polluted	Moderate
Low	None	Adaptive	Decreased	Medium	Pristine	Severe
High	Nonsense	Innate	No change	Medium	Semi-polluted	Moderate
Low	Nonsense	Adaptive	Increased	High	Polluted	Moderate
Low	Frameshift	Adaptive	No change	High	Polluted	Severe

Numeričke opisne metode

- Učestalosti (brojevi)
- Relativne frekvencije (postoci, proporcije)
- Kumulativne frekvencije
- Kumulativne relativne frekvencije
- Unakrsne (kontingencijske) tablice

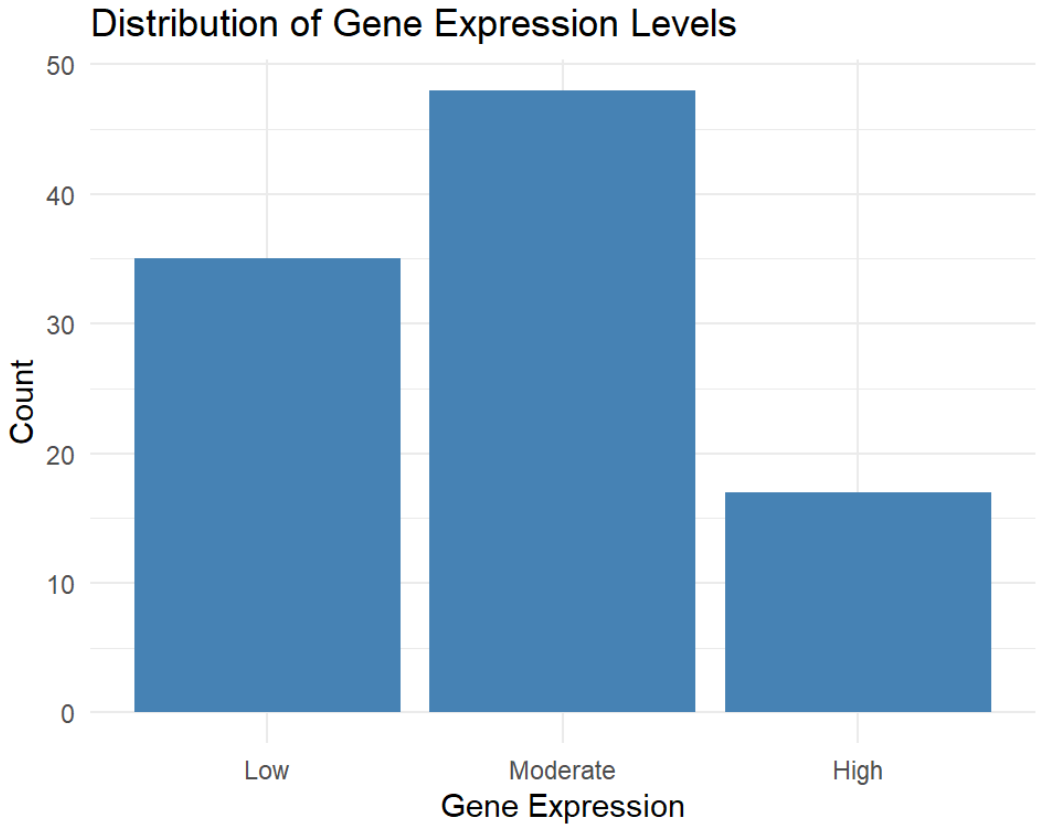
GeneExpression	Freq
Low	35
Moderate	48
High	17

GeneExpression	Proportion
Low	0.35
Moderate	0.48
High	0.17

		Activity Level		
		Increased	No change	Decreased
Gene expression	Low	11	13	11
	Moderate	18	15	15
	High	8	4	5

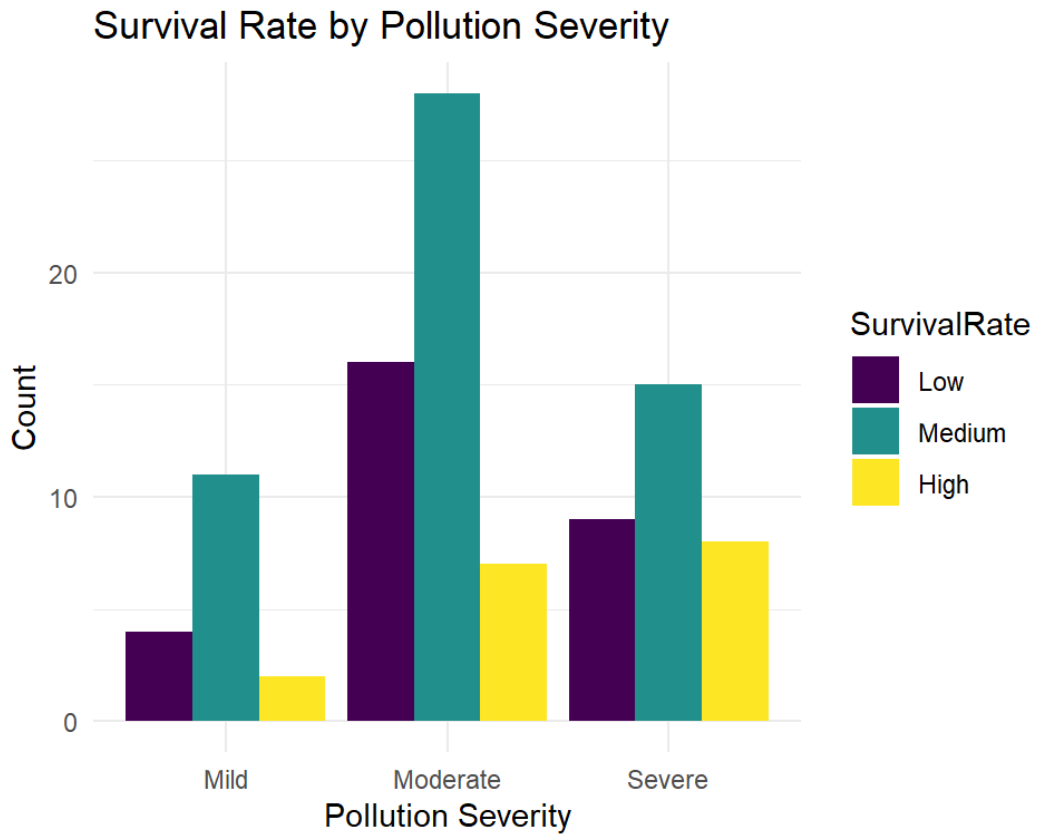
Grafičke opisne metode - Stupčasti dijagram (bar plot)

```
ggplot(data, aes(x = GeneExpression)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Distribution of Gene Expression Levels", x = "Gene Expression", y = "Count") +
  theme_minimal()
```



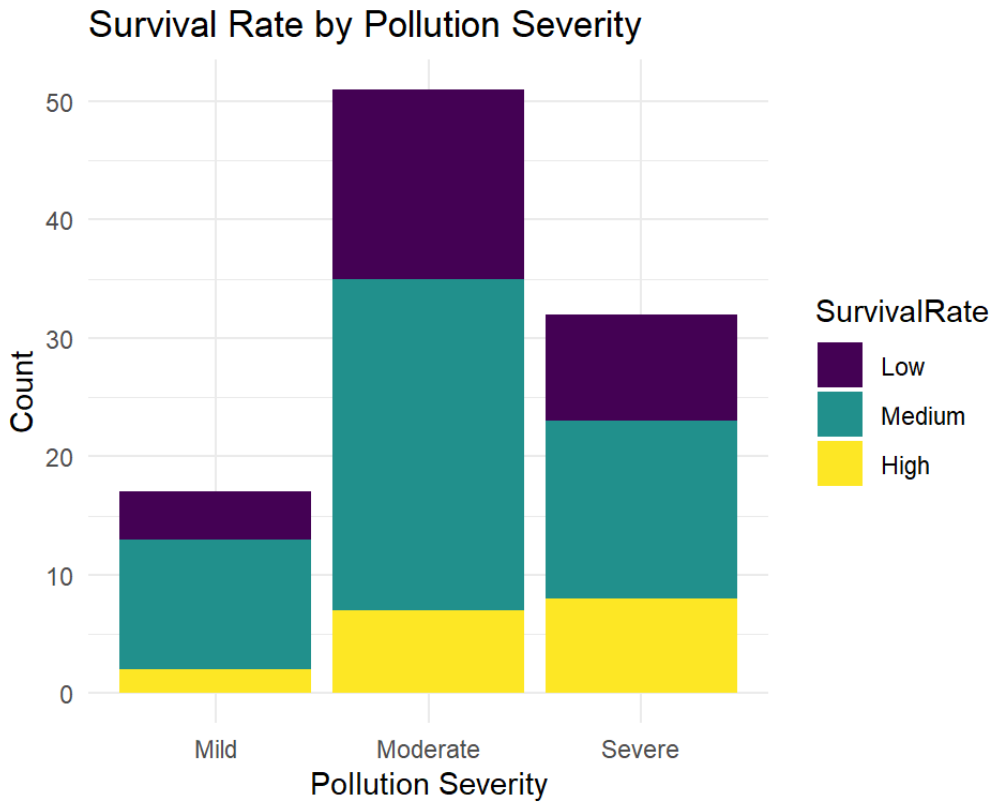
Stupčasti dijagram (bar plot) – više razina kategorizacije

```
ggplot(data, aes(x = PollutionSeverity, fill = SurvivalRate)) +
  geom_bar(position = "dodge") +
  labs(title = "Survival Rate by Pollution Severity", x = "Pollution Severity", y = "Count") +
  theme_minimal()
```



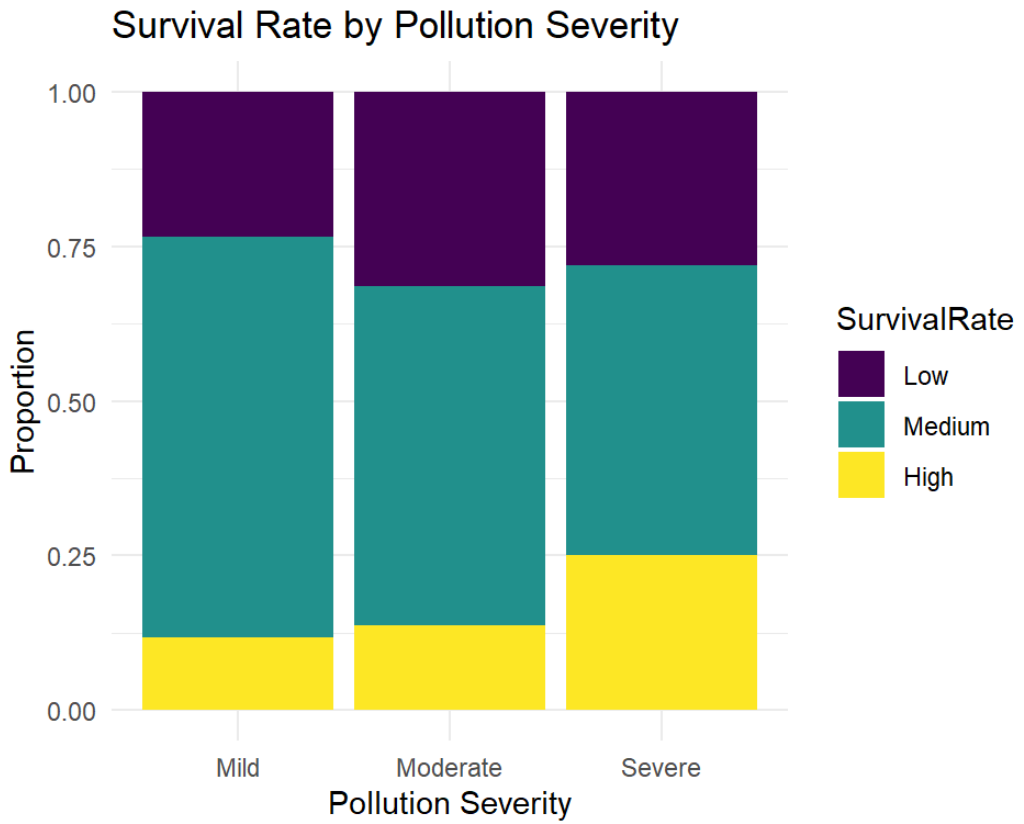
Stupčasti dijagram (bar plot) – više razina kategorizacije

```
ggplot(data, aes(x = PollutionSeverity, fill = SurvivalRate)) +
  geom_bar(position = "stack") +
  labs(title = "Survival Rate by Pollution Severity", x = "Pollution Severity", y = "Count") +
  theme_minimal()
```

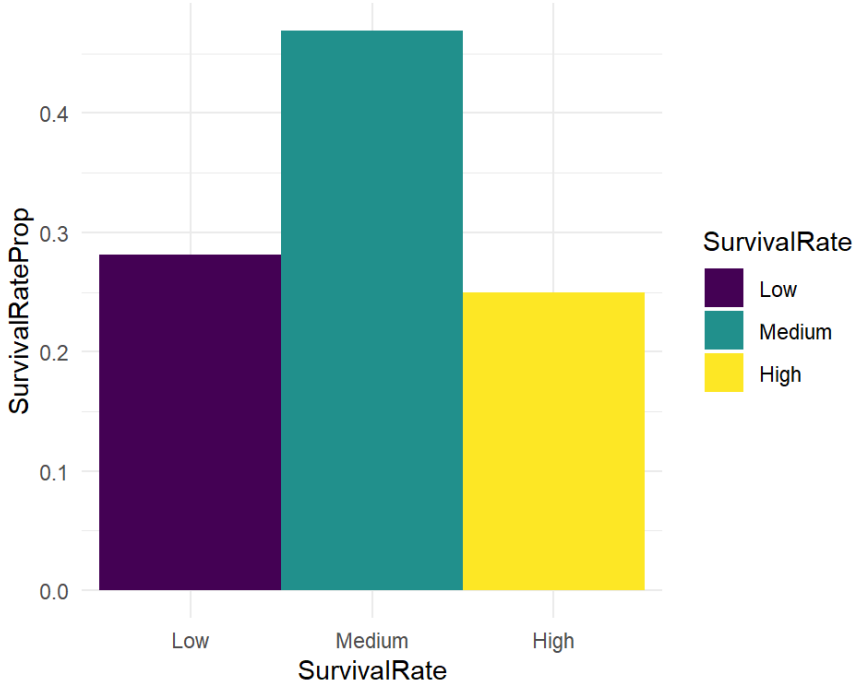


Stupčasti dijagram (bar plot) – više razina kategorizacije

```
ggplot(data, aes(x = PollutionSeverity, fill = SurvivalRate)) +
  geom_bar(position = "fill") +
  labs(title = "Survival Rate by Pollution Severity", x = "Pollution Severity", y = "Proportion") +
  theme_minimal()
```

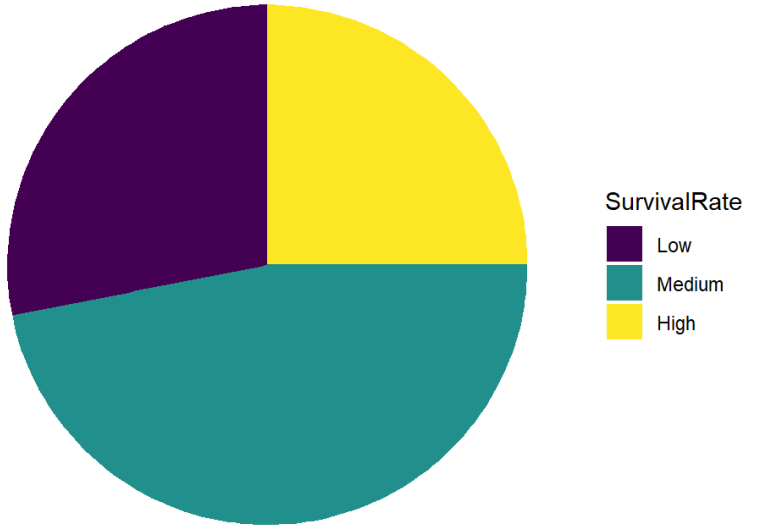


Grafičke opisne metode – Kružni dijagram (pie chart)



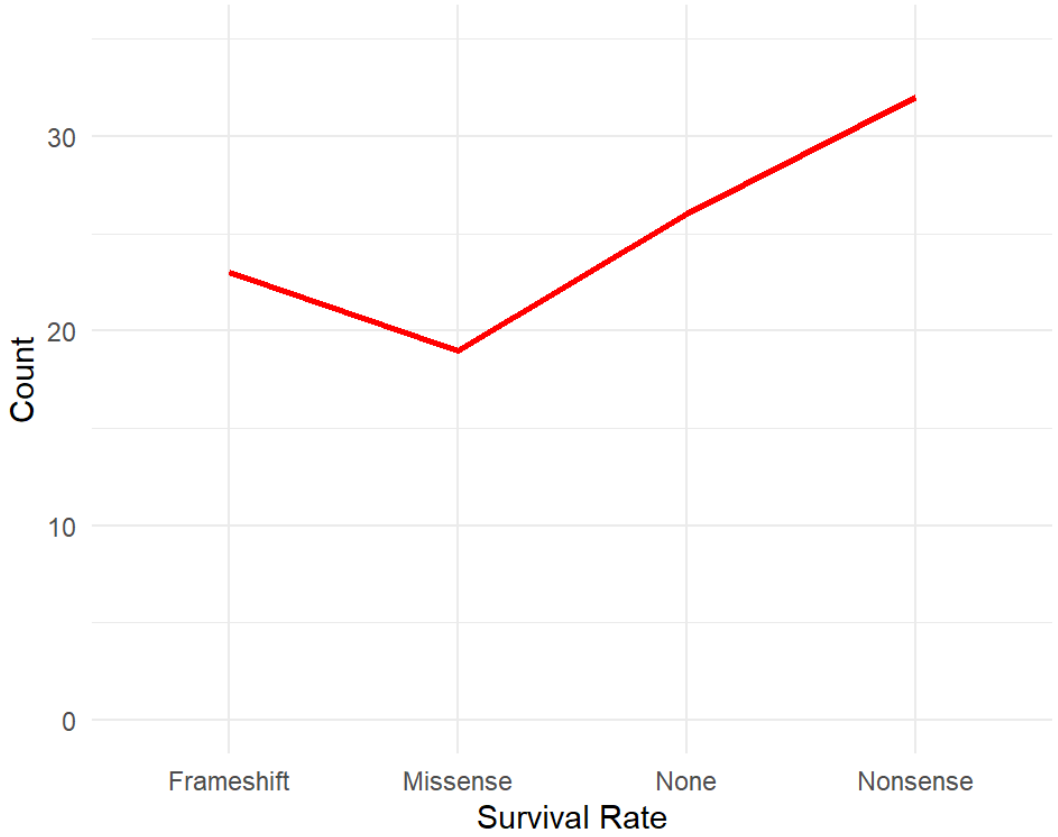
`ggplot2` does not offer any specific geom to build piecharts. The trick is the following:

- input data frame has 2 columns: the group names (`group` here) and its value (`value` here)
- build a stacked barchart with one bar only using the `geom_bar()` function.
- Make it circular with `coord_polar()`

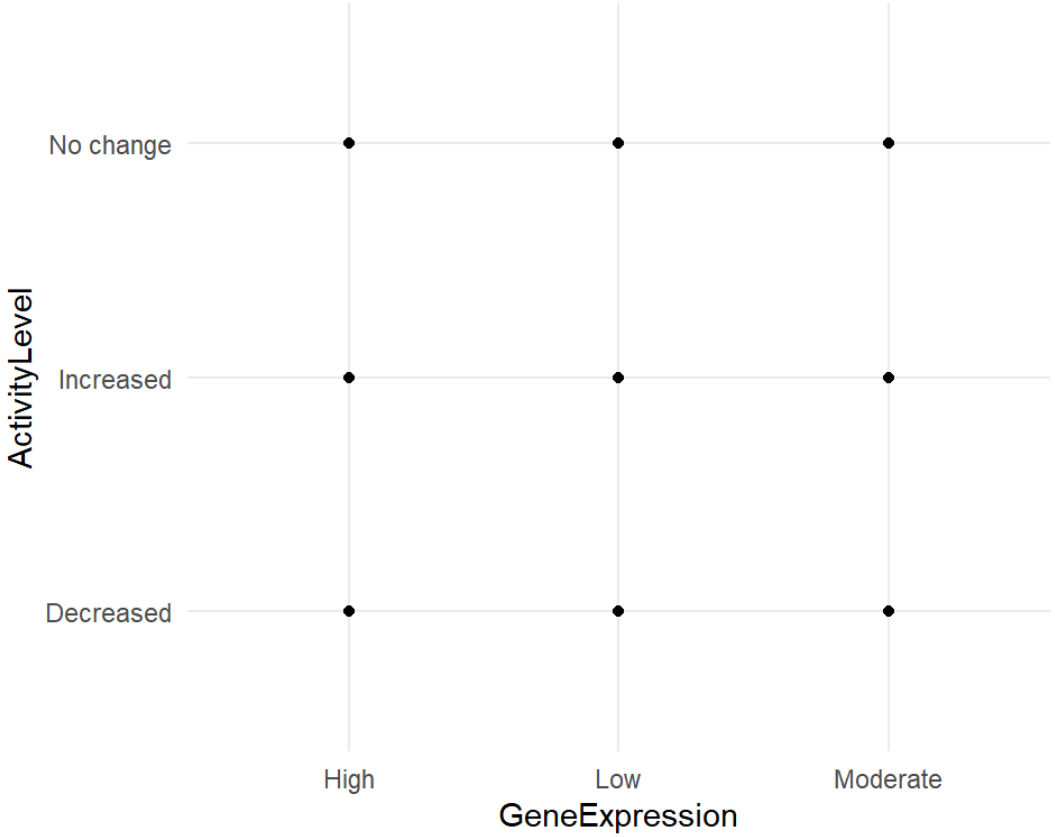


Neprimjereni grafički prikazi

Linijski graf



Dijagram raspršenja (scatterplot)



Binomni test

- Testira razlikuje li se promatrani udio uspjeha (ili jedne kategorije) značajno od očekivanog udjela
- Pretpostavke:
  - Binarni podaci (dvije kategorije: uspjeh/neuspjeh ili da/ne).
  - Neovisni eksperimenti

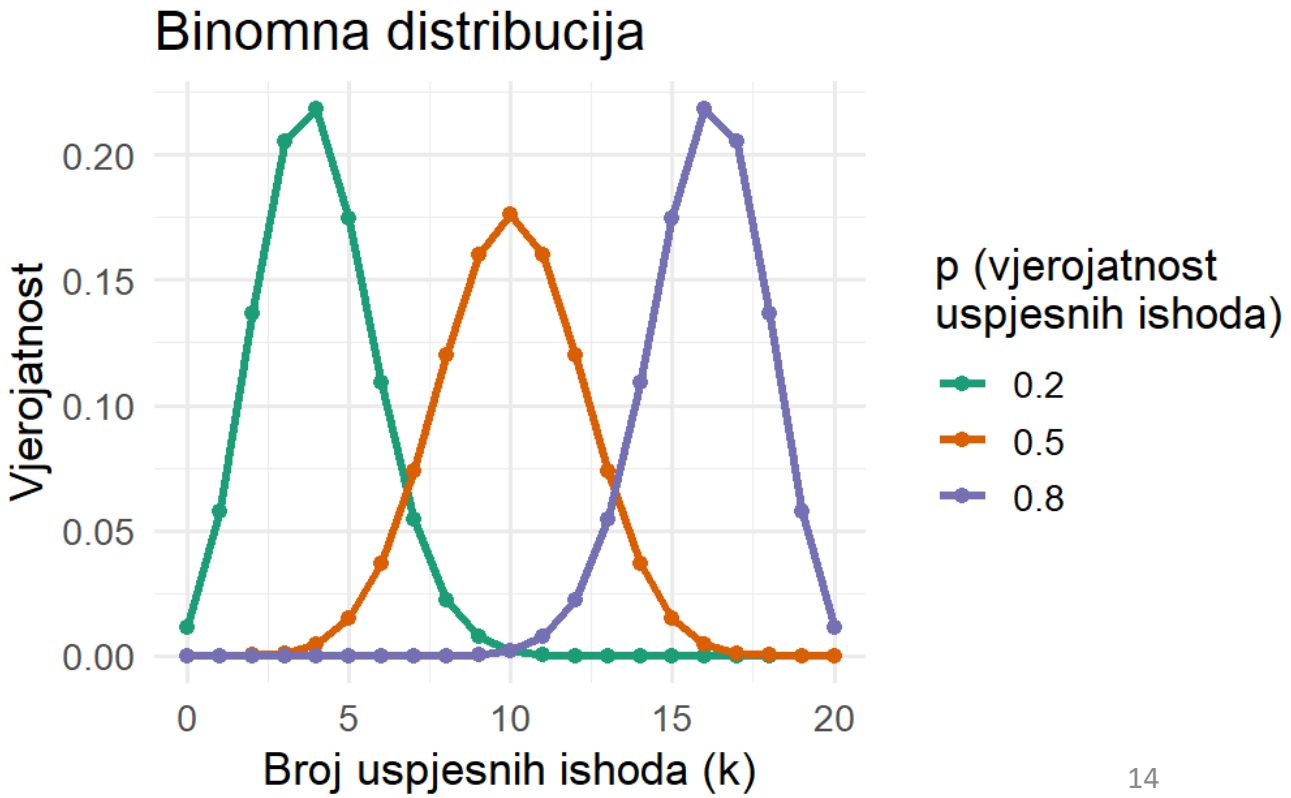
- Uspjeh - jedan od dva moguća ishoda u eksperimentu
- Definicija uspjeha specifična je za studiju koja se provodi i ovisi o istraživačkom pitanju (ishod koji proučavamo)
  - Biljka koja preživljava sušu.
  - Pacijent koji se oporavio nakon tretmana.
  - Životinja koja pokazuje specifično ponašanje (npr. preferencija za određeni okoliš)

Binomni test

- Nulta hipoteza (H_0): promatrani udio uspješnih ishoda odgovara očekivanom udjelu (p_0).
- p-vrijednost se računa na temelju binomne distribucije
- Koliko su ekstremni opaženi podaci ako je nulta hipoteza točna?

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

n – broj pokušaja (eksperimenata)
k –broj uspješnih ishoda
p – očekivani udio uspješnih ishoda



Pokazuju li vodozemci preferenciju za nezagađeni okoliš?

- Vodozemci se smještaju u odabranu komoru s dva odjeljka: jedan je zagađen PCB-ima, a drugi je nezagađen.

- Istraživači žele utvrditi preferiraju li vodozemci jedno okruženje u odnosu na drugo.
  - Nulta hipoteza (H0): Vodozemci nemaju preferencije i provodit će jednako vrijeme u oba dijela (p=0,5).
  - Alternativna hipoteza (Ha): Vodozemci preferiraju jedno okruženje u odnosu na drugo (p≠0,5)

- Rezultati: Testiramo 30 vodozemaca, 20 provodi više vremena u nezagađenom okolišu

Binomni test – jednostrani (one-tailed)

```r
# Observed data
unpolluted_preference <- 20
total_amphibians <- 30

# Binomial test
binom.test(x = unpolluted_preference,
           n = total_amphibians,
           p = 0.5,
           alternative = "g")
```

```
##
##  Exact binomial test
##
## data:  unpolluted_preference and total_amphibians
## number of successes = 20, number of trials = 30, p-value = 0.04937
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.5005613 1.0000000
## sample estimates:
## probability of success
##              0.6666667
```

Odbacujemo nultu hipotezu na razini značajnosti 0.05!
Vodozemci preferiraju nezagađeni okoliš.

Binomni test – dvostrani (two-tailed)

```
# Binomial test
binom.test(x = unpolluted_preference,
           n = total_amphibians,
           p = 0.5)
```

```
##
##  Exact binomial test
##
## data:  unpolluted_preference and total_amphibians
## number of successes = 20, number of trials = 30, p-value = 0.09874
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4718800 0.8271258
## sample estimates:
## probability of success
##              0.6666667
```

Ne odbacujemo nultu hipotezu na razini značajnosti 0.05!
Vodozemci ne pokazuju preferenciju za okoliš.

Hi-kvadrat test hipoteze (*goodness-of-fit*)

- Primjer:
- Hipoteza: Tri vrste žaba (gatalinka, zelena žaba, močvarna žaba) ravnomjerno su raspoređene u staništu
- Uzimamo slučajni uzorak od 300 žaba iz staništa i dobivamo 89 gatalinki, 120 zelenih žaba i 91 močvarnih žaba
- Je li narušena ekologija staništa?

	Gatalinka	Zelena žaba	Močvarna žaba	TOTAL
Opaženo	89 (29.7%)	120 (40.0%)	91 (30.3%)	300
Očekivano	100 (33.33 %)	100 (33.33 %)	100 (33.33 %)	300

Hi-kvadrat test (*goodness-of-fit*)

- Koliko se dvije distribucije prosječno razlikuju jedna od druge

$$\frac{\text{observed frequency} - \text{expected frequency}}{\text{expected frequency}}$$

gatalinka: $\frac{(89-100)}{100} = -0.11$ zelena žaba: $\frac{(120-100)}{100} = +0.20$ močvarna žaba: $\frac{(91-100)}{100} = -0.09$

$$\frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

gatalinka: $\frac{(89-100)^2}{100} = 1.21$ zelena žaba: $\frac{(120-100)^2}{100} = 4$ močvarna žaba: $\frac{(91-100)^2}{100} = 0.81$

zbroj: 6.02

Hi-kvadrat (χ^2) statistika

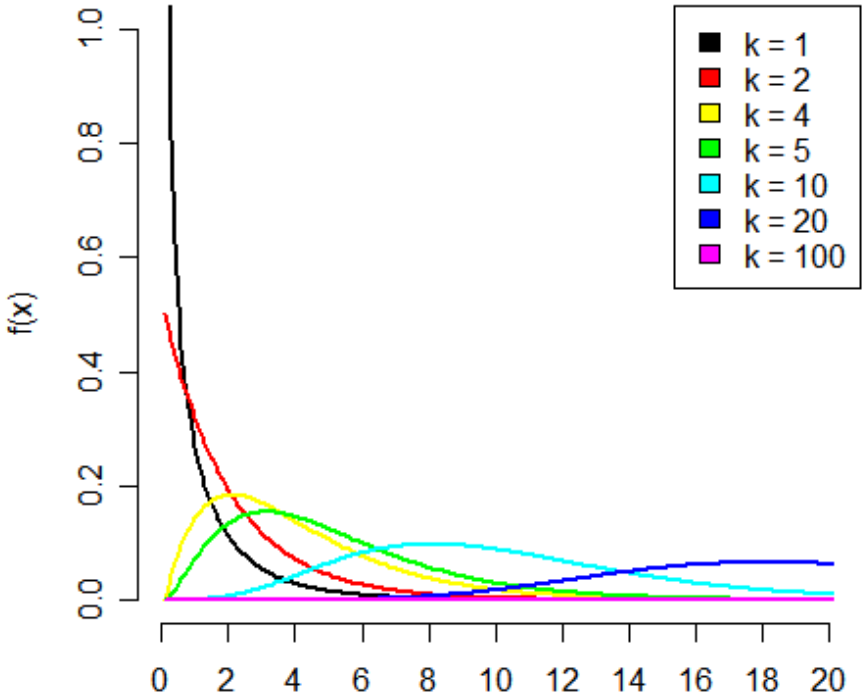
$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

- O_i – opažena frekvencija (observed frequency)
- E_i – očekivana frekvencija (expected frequency)
- N – broj kategorija

Hi-kvadrat distribucija

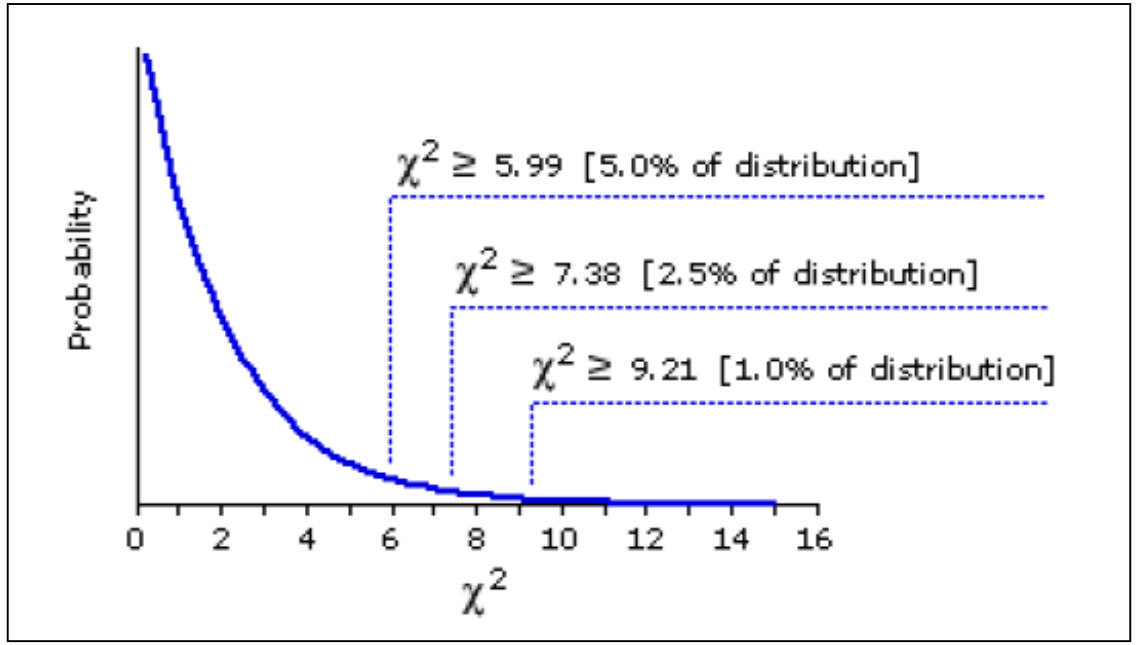
- **hi-kvadrat distribucija s *k* stupnjeva slobode je distribucija zbroja kvadrata *k* nezavisnih standardnih normalnih slučajnih varijabli**

$$f_x(X) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2-1)} e^{-x/2}$$



- U slučaju jednodimenzionalne klasifikacije broj stupnjeva slobode = broj klasa (ćelija) -1

Hi-kvadrat test



$P(X \geq 6.02) = 0.0492$

Hi-kvadrat test je neusmjeren!

Hi-kvadrat za dvije dimenzije kategorizacije

- Tablica kontingencije (unakrsna tablica) - koristi se za bilježenje i analizu odnosa između dvije ili više kategoričkih varijabli
- Prikazuje (multivarijantnu) distribuciju frekvencija varijabli u matričnom formatu
- Koristi se za testiranje povezanosti (neovisnosti) između dvije ili više razina klasifikacije

		Visoka razina aktivnosti		
		DA	NE	
Visoka ekspresija gena	DA	32	18	50
	NE	23	77	100
		55	95	150

Hi-kvadrat za dvije dimenzije kategorizacije

- Želimo utvrditi jesu li dvije kategoričke varijable povezane – hi-kvadrat test asocijacije

		Visoka razina aktivnosti		
		DA	NE	
Visoka ekspresija gena	DA	32	18	50
	NE	23	77	100
		55	95	150

- **Primjena hi-kvadrat testa je ista osim za:**
- Izračun očekivanih frekvencija ćelija
- Izračun odgovarajuće vrijednosti za stupnjeve slobode.

Hi-kvadrat za dvije dimenzije kategorizacije

- Kako izračunati očekivane frekvencije?

		Visoka razina aktivnosti		
		DA	NE	
Visoka ekspresija gena	DA	$\frac{R1 \times C1}{TOTAL}$	$\frac{R1 \times C2}{TOTAL}$	R1
	NE	$\frac{R2 \times C1}{TOTAL}$	$\frac{R2 \times C2}{TOTAL}$	R2
		C1	C2	TOTAL

- Kako izračunati stupnjeve slobode?

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

r – broj redova

c – broj stupaca

Hi-kvadrat za dvije dimenzije kategorizacije

- Izračun hi-kvadrat statistike

  - U slučaju više od dva reda i dva stupca

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

  - U slučaju točno dva reda i dva stupca (korekcija za kontinuitet)

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

Nastavak primjera…

		Visoka razina aktivnosti		
		DA	NE	
Visoka ekspresija gena	DA	32 (E = 18.33)	18 (E = 31.67)	50
	NE	23 (E = 36.67)	77 (E = 63.33)	100
		55	95	150

- Hi-kvadrat statistika = 22.396
- Stupnjevi slobode = 1
- Vjerojatnost opažanja hi-kvadrat statistike od 22.396 ako je nulta hipoteza točna je 2.218×10^{-6}

Pretpostavke za korištenje hi-kvadrat testa

- Postupci hi-kvadrat testa mogu se legitimno primijeniti samo ako su kategorije u koje su razvrstana N opažanja neovisne jedna o drugoj.

- Logička valjanost hi-kvadrat testa je najveća kada su vrijednosti E, srednje očekivane slučajne frekvencije unutar ćelija, prilično velike, i smanjuje se kako te vrijednosti E postaju niže. Postupci hi-kvadrat testa mogu se legitimno primijeniti samo ako su sve vrijednosti E jednake ili veće od 5.

Fisherov egzaktni test

- Hi-kvadrat test može se legitimno primijeniti samo ako su sve vrijednosti E jednake ili veće od 5.
- Alternativa u slučaju tablica kontingencije 2x2 – Fisherov egzaktni test vjerojatnosti
- Izračunajte točnu vjerojatnost dobivanja promatranih vrijednosti frekvencije

  - Može se koristiti kao usmjereni ili neusmjereni test
  - Nedostatak: komplicirani izračuni

		VISOKA AKTIVNOST		
VISOKA EKSPRESIJA		NE	DA	
	DA	2	7	9
	NE	8	2	10
		10	9	19

Logika Fisherovog egzaktnog testa

- H_0: Nema povezanosti između razine ekspresije gena i razine aktivnosti
- Ako je nulta hipoteza točna koliko je vjerojatno da ćemo dobiti ovakav ili ekstremniji rezultat?

		VISOKA AKTIVNOST		
VISOKA EKSPRESIJA		NE	DA	
	DA	2	7	9
	NE	8	2	10
		10	9	19

Logika Fisherovog egzaktnog testa

		VISOKA AKTIVNOST		
VISOKA EKSPRESIJA		NE	DA	
	DA	2	7	9
	NE	8	2	10
		10	9	19

		VISOKA AKTIVNOST		
VISOKA EKSPRESIJA		NE	DA	
	DA			9
	NE			10
		10	9	19

- 10 mogućih načina dobivanja zbroja u redovima i stupcima

Ô1	Ô2	Ô3	Ô4	Ô5	Ô6	Ô7	Ô8	Ô9	Ô10
9 0	8 1	7 2	6 3	5 4	4 5	3 6	2 7	1 8	0 9
1 9	2 8	3 7	4 6	5 5	6 4	7 3	8 2	9 1	10 0

- Ako je nulta hipoteza točna, koliko je vjerojatno da ćemo dobiti ishode Ô8 or Ô9 or Ô10?

Logika Fisherovog egzaktnog testa

- Postupak Fisherovog egzaktnog testa:
  - Izračunajte vjerojatnost svakog događaja koji je "ovako velik ili veći"
  - Zbrojite disjunktne vjerojatnosti

$$\mathrm{P(outcome)} = \frac{\text{number of possibilities favourable to the occurence of the outcome}}{\text{total number of possibilities}}$$

Logika Fisherovog egzaktnog testa

- Broj mogućih kombinacija dobivanja *k* uspjeha u *N* eksperimenata:

$$\binom{N}{k} = \frac{N!}{k!\,(N-k)!}$$

- Broj mogućih načina dobivanja zbroja u redovima i stupcima za EKSPRESIJU i AKTIVNOST:

		VISOKA AKTIVNOST		
		NE	DA	
VISOKA EKSPRESIJA	DA			9
	NE			10
		10	9	19

P1-P9=DA, P10-P19=NE
P1-P8, P10=DA, P9, P11-P19=NE
itd… (92378 kombinacija)

$$\frac{19!}{9!\,(19-9)!} = 92378$$

		VISOKA AKTIVNOST		
		NE	DA	
VISOKA EKSPRESIJA	DA			9
	NE			10
		10	9	19

$$\frac{19!}{9!\,(19-9)!} = 92378$$

P1-P9=DA, P10-P19=NE
P1-P8, P10=DA, P9, P11-P19=NE
itd… (92378 kombinacija)

- Ukupan broj kombinacija = 92378 x 92378 = 8533694884

Logika Fisherovog egzaktnog testa

$$\mathrm{P(outcome)} = \frac{?}{8533694884}$$

- Kako izračunati ukupan broj kombinacija koje idu u prilog ovom ishodu?

$\hat{O}10$

		VISOKA AKTIVNOST		
VISOKA EKSPRESIJA		NE	DA	
	DA	0	9	9
	NE	10	0	10
		10	9	19

- 92,378 mogućih kombinacija gdje je 9 AKTIVNOST=DA and 10 AKTIVNOST=NE
- Za svaku kombinaciju varijable AKTIVNOST postoji točno jedna kombinacija varijable EKSPRESIJA gdje je AKTIVNOST = DA ⟶ EKSPRESIJA = DA and ESPRESIJA = NE ⟶ AKTIVNOST = NE

$$\mathrm{P}(\hat{O}10) = \frac{92378}{8533694884} = \frac{1}{92378}$$

Logika Fisherovog egzaktnog testa

- 92,378 mogućih kombinacija za 9 AKTIVNOST=DA and 10 AKTIVNOST=NE
- Broj kombinacija koje će proizvesti odgovarajuću raspodjelu EKSPRESIJE i

  AKTIVNOSTI u stupcu AKTIVNOST = NE je $\binom{10}{2} = 45$
- Broj kombinacija koje će proizvesti odgovarajuću raspodjelu EKSPRESIJE i

  AKTIVNOSTI u stupcu AKTIVNOST = DA $\binom{9}{7} = 36$

$\hat{O}8$

		VISOKA AKTIVNOST		
VISOKA EKSPRESIJA		NE	DA	
	DA	2	7	9
	NE	8	2	10
		10	9	19

- Broj kombinacija kojim se mogu proizvesti ukupne frekvencije:

  45 x 36 x 92378 = 149652360

$$P(\hat{O}8) = \frac{149652360}{8533694884} = \frac{1620}{92378}$$

Logika Fisherovog egzaktnog testa

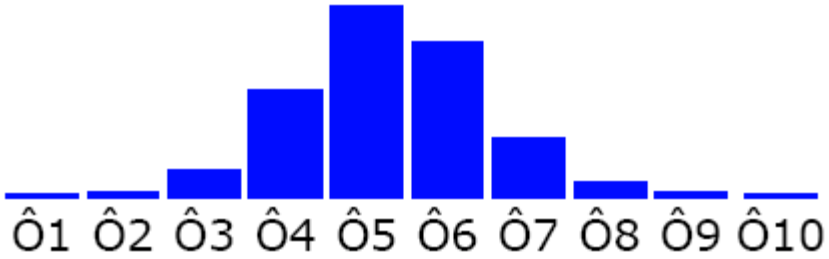
- Ukupna vjerojatnost dobivanja opaženih ili većih (ekstremnijih) frekvencija:

$$P(\hat{O}8) + P(\hat{O}9) + P(\hat{O}10) =$$

$$1620/92378 + 90/92378 + 1/92378 = 0.0185$$

Fisherov egzaktni test za neusmjerenu hipotezu

- Distribucija rezultata uzorkovanja:



- Usmjerena hipoteza: Promatrajte ishod "ovako velik ili veći"
- Neusmjerena hipoteza: provjerite postoji li veza u oba smjera (promatrajte ishod "ovako velik ili veći" u oba repa distribucije)

Fisherov egzaktni test za neusmjerenu hipotezu

- Koncept "ovako velik ili veći" definiran je stupnjem do kojeg su frekvencije unutar ćelija raspoređene neproporcionalno
- Stupanj povezanosti unutar niza frekvencija unutar ćelija mjeri se pomoću mjere neproporcionalnosti (*disproportion*):

a	b
c	d

$$\text{disproportion} = \left| \frac{a}{a+b} - \frac{c}{c+d} \right|$$

Ô1	Ô2	Ô3	Ô4	Ô5	Ô6	Ô7	Ô8	Ô9	Ô10
9 0	8 1	7 2	6 3	5 4	4 5	3 6	2 7	1 8	0 9
1 9	2 8	3 7	4 6	5 5	6 4	7 3	8 2	9 1	10 0
0.90	**0.69**	0.48	0.27	0.06	0.16	0.37	**0.58**	**0.79**	**1.00**

Fisherov egzaktni test za neusmjerenu hipotezu

- Za neusmjerenu hipotezu zbrajamo vjerojatnosti ishoda za koje je nesrazmjer "jednako velik ili veći" kao i nesrazmjer ishoda koji proučavamo

- Neusmjerena dvostrana vjerojatnost:

$$P(\hat{O}1) + P(\hat{O}2) + P(\hat{O}8) + P(\hat{O}9) + P(\hat{O}10) = 0.023$$

Ostale metode za kategoričke podatke

- **McNemarov test** je statistički test koji se koristi za analizu uparenih kategoričkih podataka, posebno za 2x2 tablice kontingencije. Testira postoji li značajna promjena u omjerima između dvije povezane skupine ili stanja.
- **Logistička regresija** - modeliranje odnosa između jedne ili više nezavisnih varijabli (prediktora) i binarne ovisne varijable (ishod).

Primjer 1.

Želite istražiti razlikuje li se distribucija ponašanja vodozemaca (pojačana aktivnost, smanjena aktivnost, bez promjene) u tri tipa staništa (zagađeno, poluzagađeno, netaknuto).

Koji test trebate koristiti?

(a) Hi-kvadrat test neovisnosti

(b) t-test

(c) Fisherov egzaktni test

Primjer 2.

Testiranje stopa preživljavanja i ozbiljnosti onečišćenja

Pretpostavljate da ozbiljnost onečišćenja (blaga, umjerena, jaka) utječe na kategoriju stope preživljavanja (niska, srednja, visoka).

Koji test trebate koristiti?

(a) Hi-kvadrat test hipoteze (goodness-of-fit)

(b) Hi-kvadrat test neovisnosti

(c) Logistička regresija

Primjer 3.

Usporedba vrsta imunološkog odgovora

Želite utvrditi je li vrsta imunološkog odgovora (urođena, adaptivna) jednako raspoređena u skupu podataka.

Koji test trebate koristiti?

(a) Binomni test

(b) t-test jednog uzorka

(c) Hi-kvadrat test hipoteze (goodness-of-fit)