Testiranje razlika u srednjoj vrijednosti (parametarijske i neparametrijske metode)

03.12.2024.

Rosa Karlić

Analize bioloških podataka 2024/2025

Zašto testirati razlike u srednjoj vrijednosti?

Primjer:

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.

Zašto testirati razlike u srednjoj vrijednosti?

- **Testiranje znanstvenih ili istraživačkih hipoteza o parametrima populacije.**

  Povećava li PCB razinu ekspresije CYP1A1?

- **Jesu li uočene razlike posljedica prirodne varijacije ili značajnog učinka.**

  Jesu li razlike razini ekspresije CYP1A1 uzrokovane tretmanom ili slučajnom varijacijom?

- **Kvantifikacija veličine učinka (ne samo postoji li razlika, nego i kolika je razlika)**

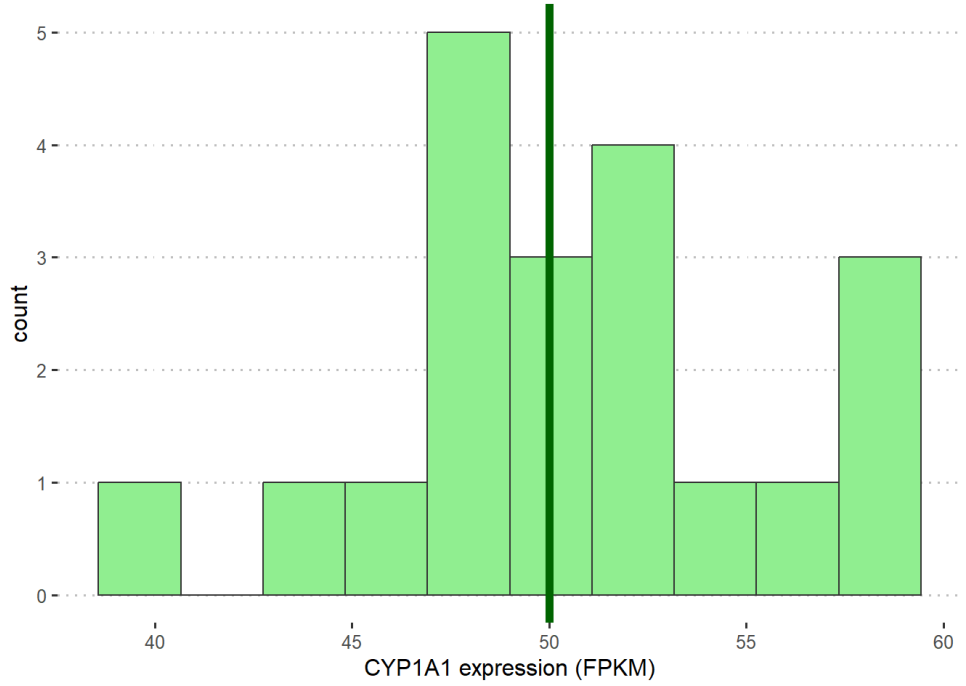
  Koliko razina PCB-a utječe ekspresiju gena CYP1A1?

- **Identifikacija trendova tijekom vremena (promjene unutar iste grupe ispitanika/subjekata)**

  Je li se razina ekspresije gena CYP1A1 vratila na normalnu razinu nakon uklanjanja onečišćenja?

Testiranje hipoteza

- Na temelju prethodnih istraživanja pretpostavljamo da je srednja vrijednost ekspresije gena CYP1A1 50 FPKM.
- Kako testirati ovu hipotezu?
- Podupiru li naši podaci ovu hipotezu?



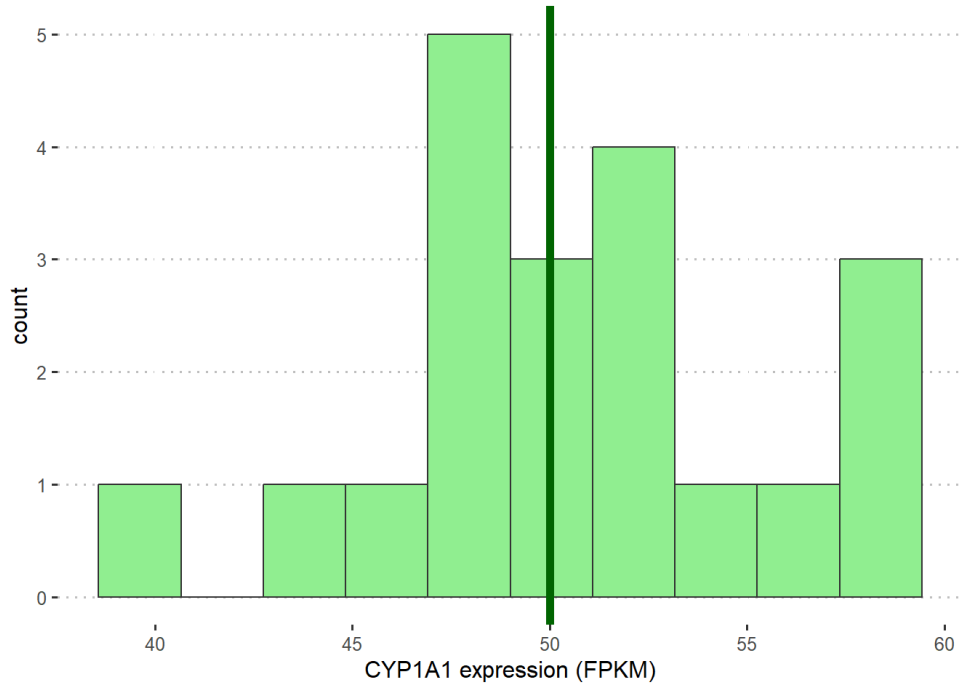
Testiranje statističkih hipoteza

- **Statistička hipoteza –** pretpostavka o parametru populacije (može i ne mora biti istinita)
- **Nulta hipoteza** (H_0) – hipoteza koju testiramo, uglavnom hipoteza koja govori da su naša opažanja rezultat slučajnosti
- **Alternativna hipoteza**(H_1 ili H_a) – hipoteza koja govori da na ispitanike u uzorku utječe neki ne-slučajni (non-random) uzrok

- Ispitujemo nasumičan uzorak iz populacije
- Ako su podaci iz uzorka u skladu s nultom hipotezom (na određenoj razini pouzdanosti) ne odbacujemo ju, u suprotnom – odbacujemo nultu hipotezu

Testiranje hipoteza

- Na temelju prethodnih istraživanja pretpostavljamo da je srednja vrijednost ekspresije gena CYP1A1 50 FPKM.
- Kako testirati ovu hipotezu?
- Podupiru li naši podaci ovu hipotezu?



$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

$$\text{test statistika} = \frac{\text{sistematska varijacija}}{\text{nesistematska varijacija}} = \frac{\text{učinak}}{\text{greška}}$$

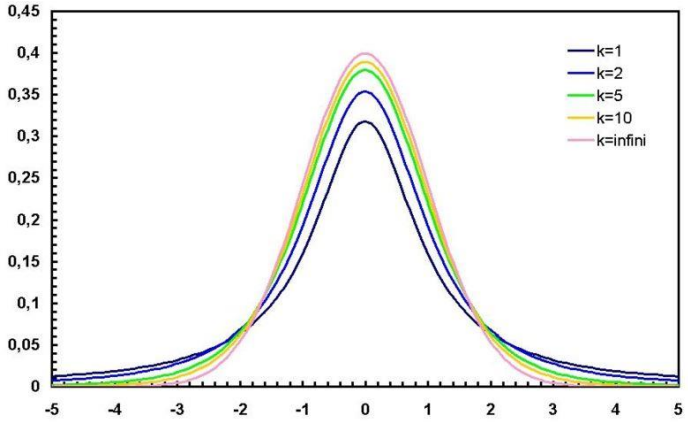
Studentov t-test (jedan uzorak)

- t-statistika

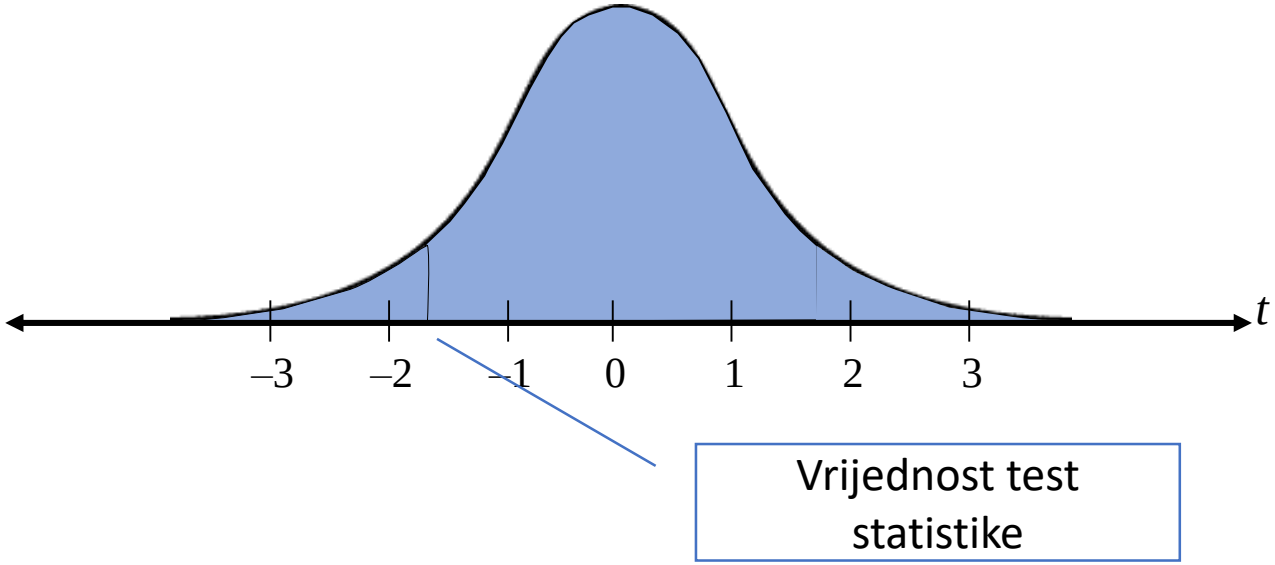
$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

$$SE_{\bar{x}} = s / \sqrt{n}$$

- Ukoliko je nulta hipoteza točna t-statistika će slijediti t-distribuciju sa n-1 stupnjeva slobode



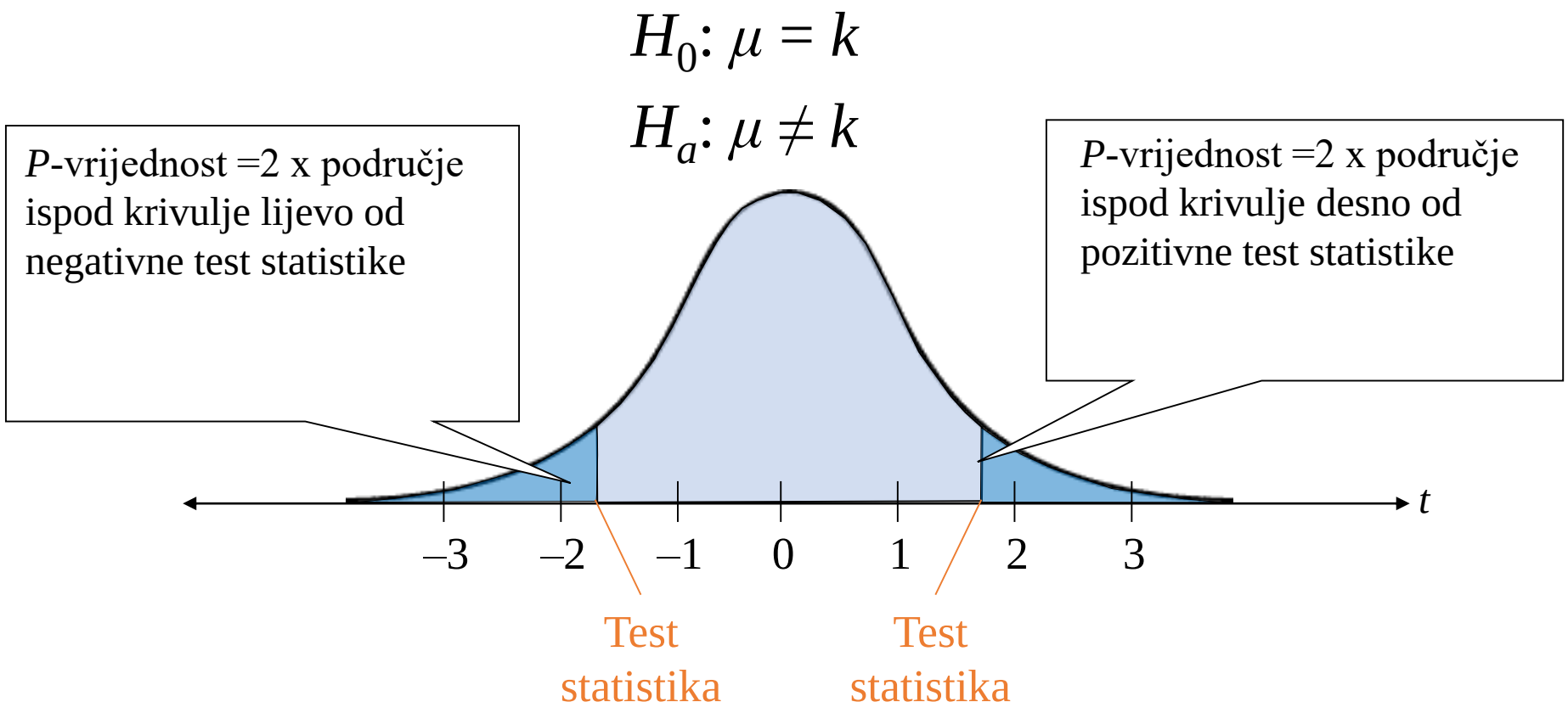
$$f_\nu(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-(\nu+1)/2}$$



Studentov t-test (jedan uzorak)

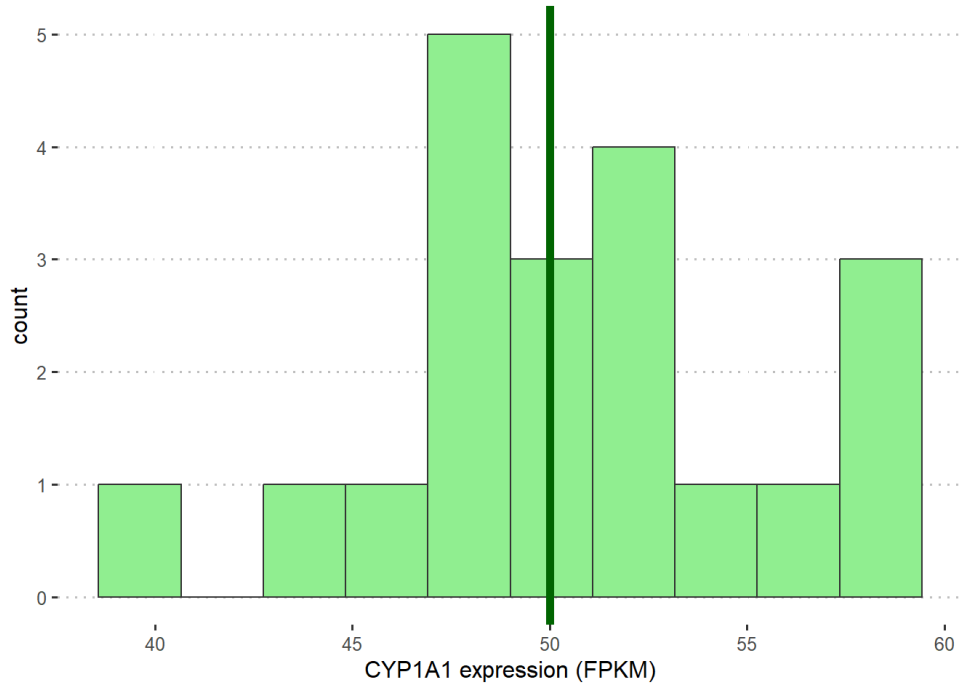
$H_0: \mu = k$

$H_a: \mu \neq k$



Testiranje hipoteza

- Na temelju prethodnih istraživanja pretpostavljamo da je srednja vrijednost ekspresije gena CYP1A1 50 FPKM.
- Kako testirati ovu hipotezu?
- Podupiru li naši podaci ovu hipotezu?



One-sample t-test

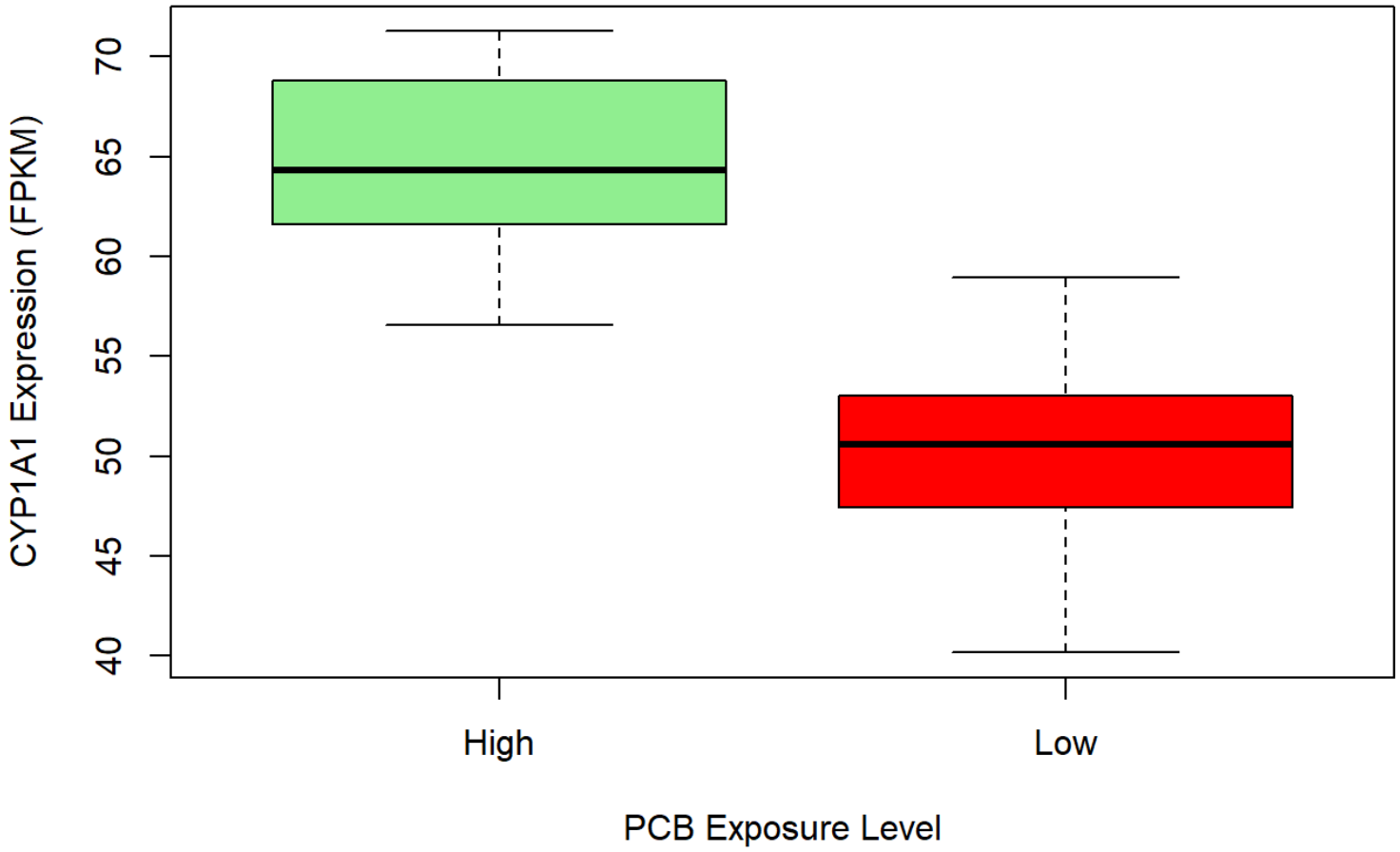
```
t.test(pcb_data[pcb_data$Exposure == "Low", "GeneExpression"], mu
= 50)
```

```
##
##  One Sample t-test
##
## data:  pcb_data[pcb_data$Exposure == "Low", "GeneExpression"]
## t = 0.65116, df = 19, p-value = 0.5227
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  48.43201 52.98423
## sample estimates:
## mean of x
##  50.70812
```

Ne odbacujemo nultu hipotezu!

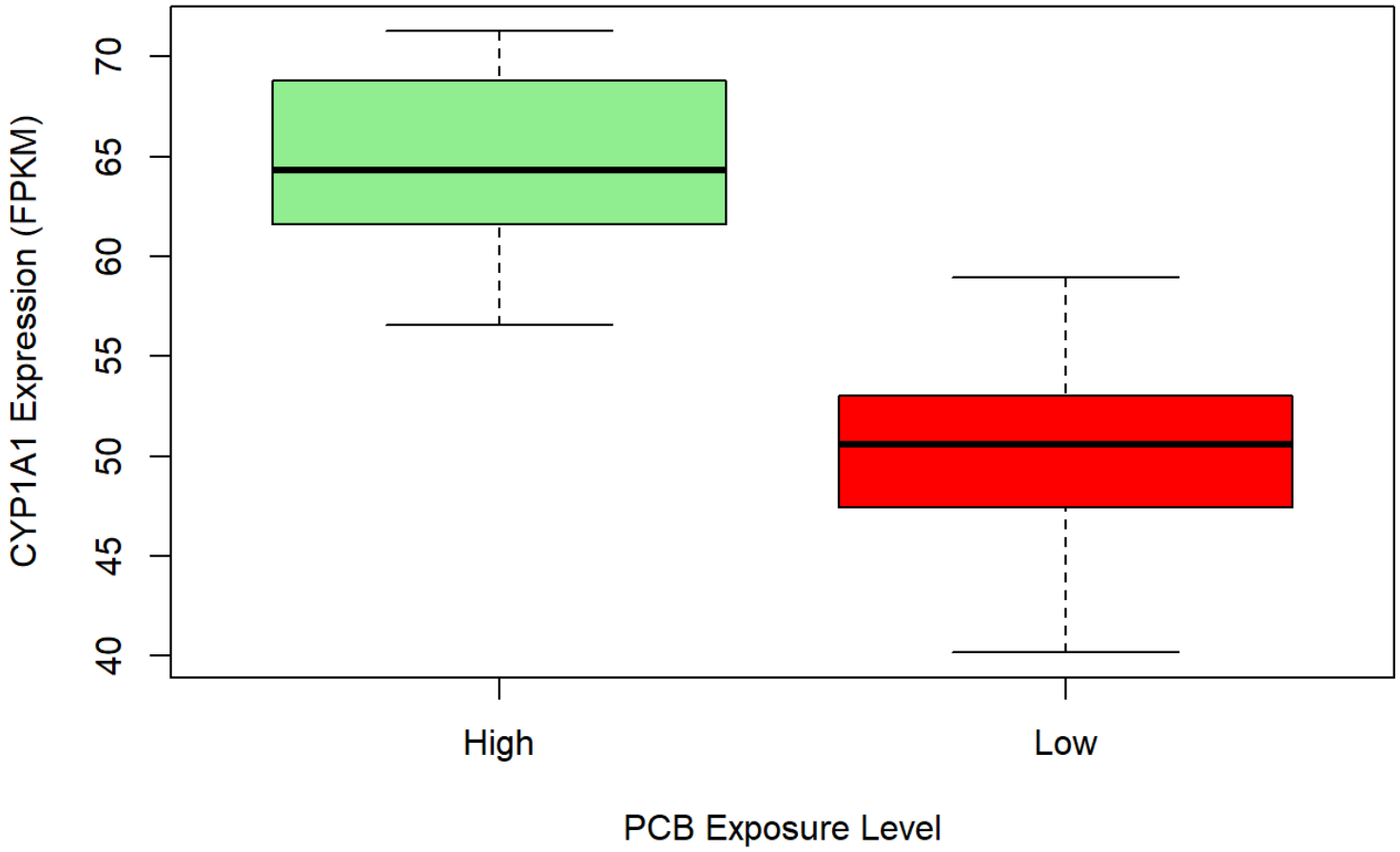
Što ako želimo usporediti dvije različite skupine?

- Ekspresija gena CYP1A1 u kontrolnoj skupini i u vodozemaca izloženih PCB-u.



Što ako želimo usporediti dvije različite skupine?

- Ekspresija gena CYP1A1 u kontrolnoj skupini i u vodozemaca izloženih PCB-u.



$$t_{n_1+n_2-2} = \frac{\bar{X}_1 - \bar{X}_2}{SE_{\bar{X}_1-\bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_P{}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Studentov t-test (dva uzorka)

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

```
# Perform t-test to compare gene expression between low and high PCB exposure
t_test_result <- t.test(GeneExpression ~ Exposure, data = pcb_data, var.equal = T )
print(t_test_result)
```

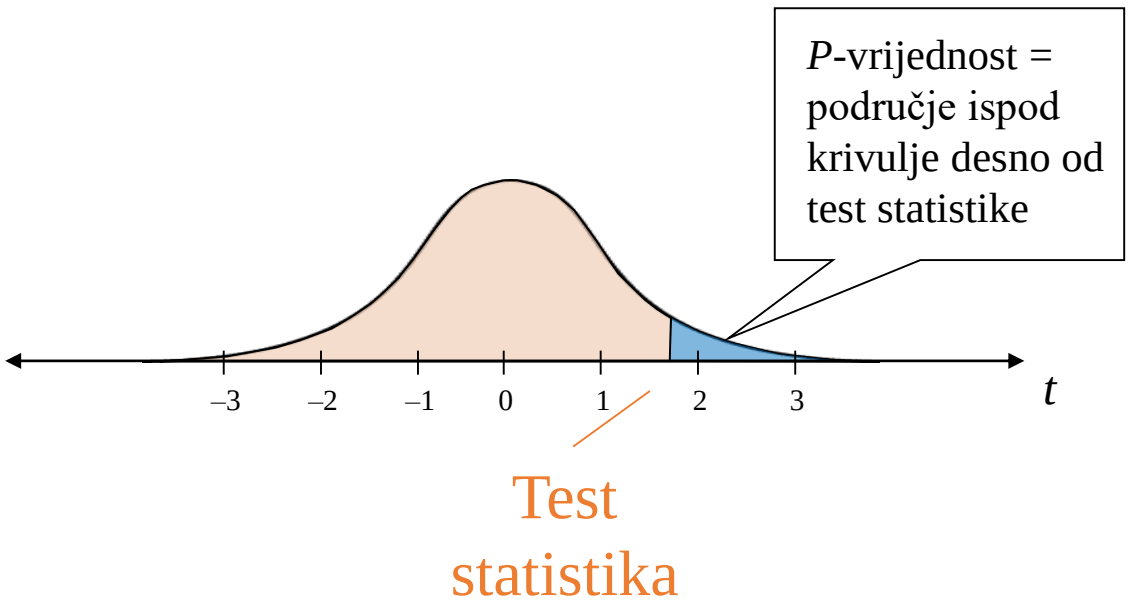
```
##
##  Two Sample t-test
##
## data:  GeneExpression by Exposure
## t = 9.8182, df = 38, p-value = 5.67e-12
## alternative hypothesis: true difference in means between group High and group Low is not
equal to 0
## 95 percent confidence interval:
##  11.14164 16.92955
## sample estimates:
## mean in group High  mean in group Low
##           64.74371           50.70812
```

Odbacujemo nultu hipotezu na razini značajnosti 0.05!

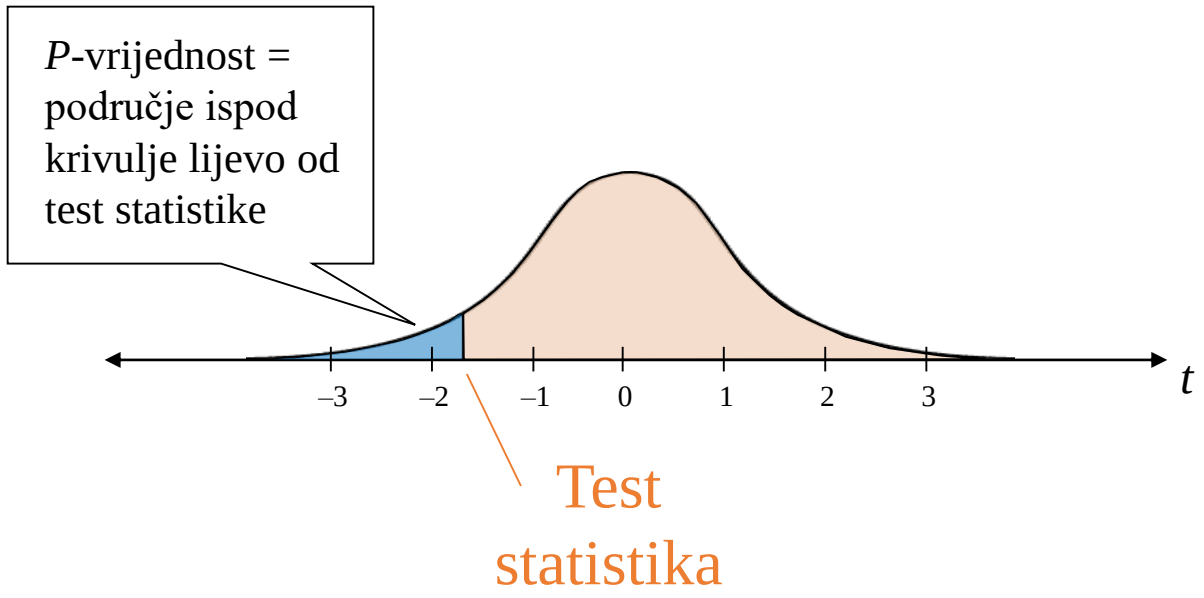
Jednostrani (one-tailed) Studentov t-test (dva uzorka)

- Nas zapravo zanima povećava li se ekspresija gena CYP1A1 kod vodozemaca izloženih PCB-u.

$$H_0: \mu \leq k$$
$$H_a: \mu > k$$



$$H_0: \mu \geq k$$
$$H_a: \mu < k$$



Jednostrani (one-tailed) Studentov t-test (dva uzorka)

- Nas zapravo zanima povećava li se ekspresija gena CYP1A1 kod vodozemaca izloženih PCB-u.

```
# Perform t-test to compare gene expression between low and high PCB exposure
t_test_result_one <- t.test(GeneExpression ~ Exposure, data = pcb_data,  var.equal = T,
alternative = "g")
print(t_test_result_one)
```

```
##
##  Two Sample t-test
##
## data:  GeneExpression by Exposure
## t = 9.8182, df = 38, p-value = 2.835e-12
## alternative hypothesis: true difference in means between group High and group Low is
greater than 0
## 95 percent confidence interval:
##  11.62545      Inf
## sample estimates:
## mean in group High  mean in group Low
##           64.74371           50.70812
```

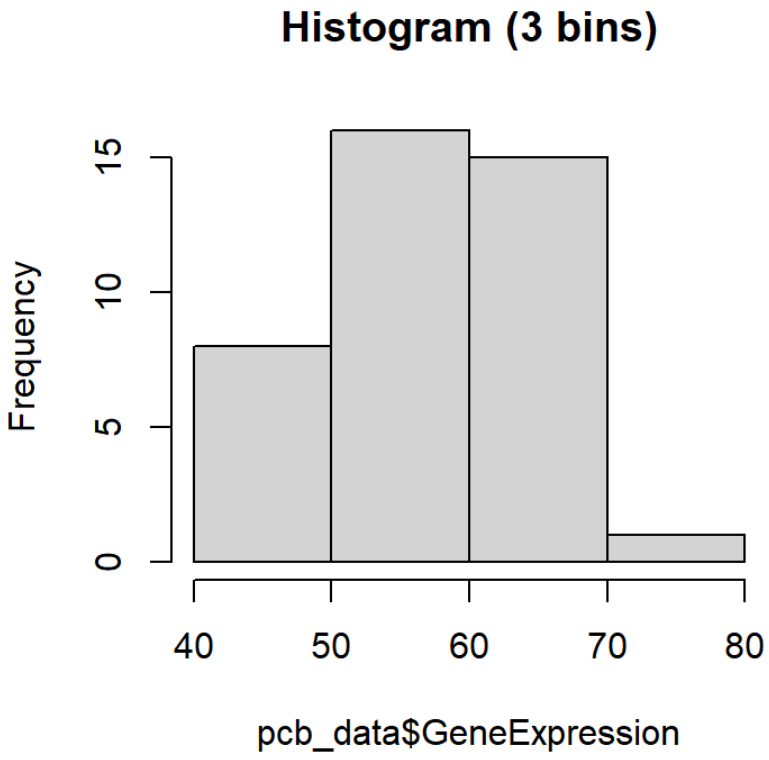
Pretpostavke za t-test

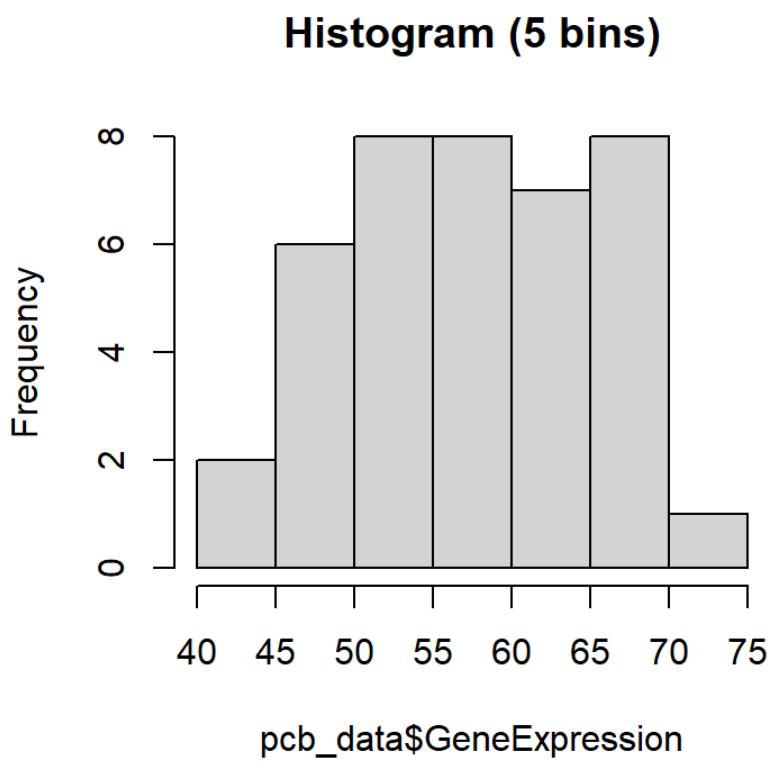
- Varijabla koju testiramo je normalno distribuirana
- Uzorci imaju jednake varijance
- Uzorci su međusobno neovisni

Testiranje normalnosti:

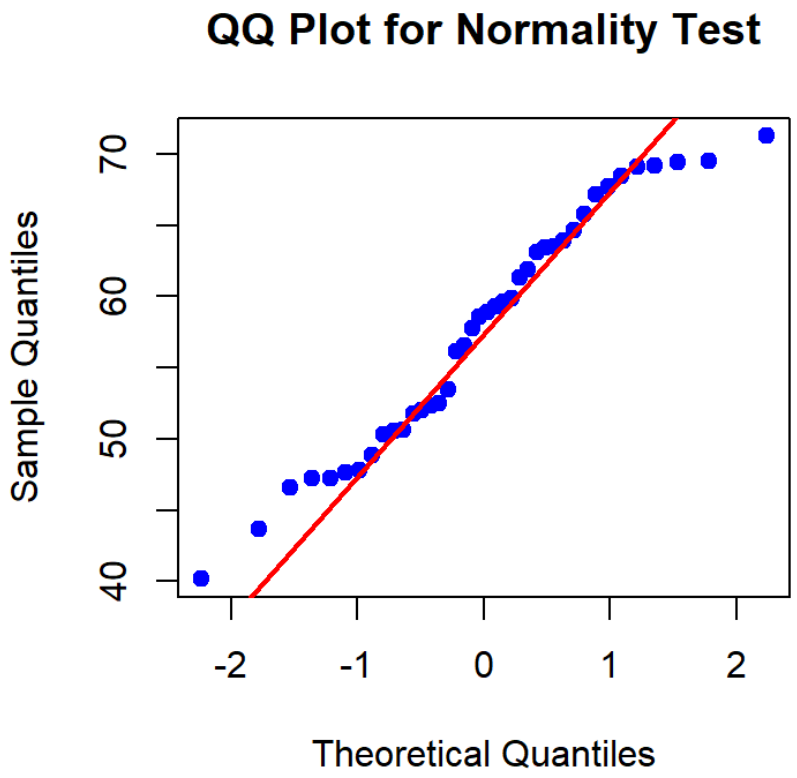
- Histogram, QQplot (kvantil-kvantil graf)
- Statistički testovi (Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling) – mogu biti preosjetljivi ako je n > 50

Histogram za testiranje normalnosti



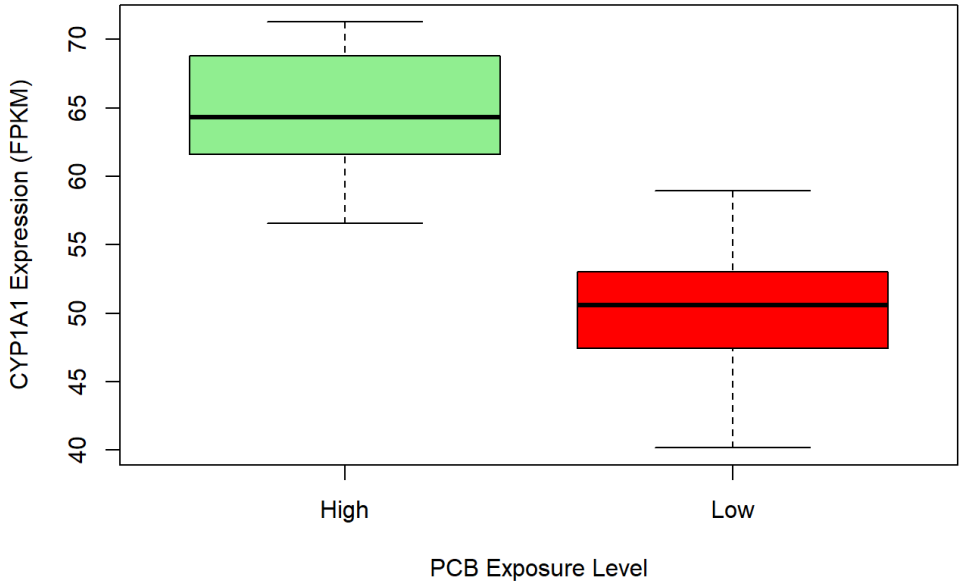


QQplot za testiranje normalnosti



Testiranje homogenosti varijance

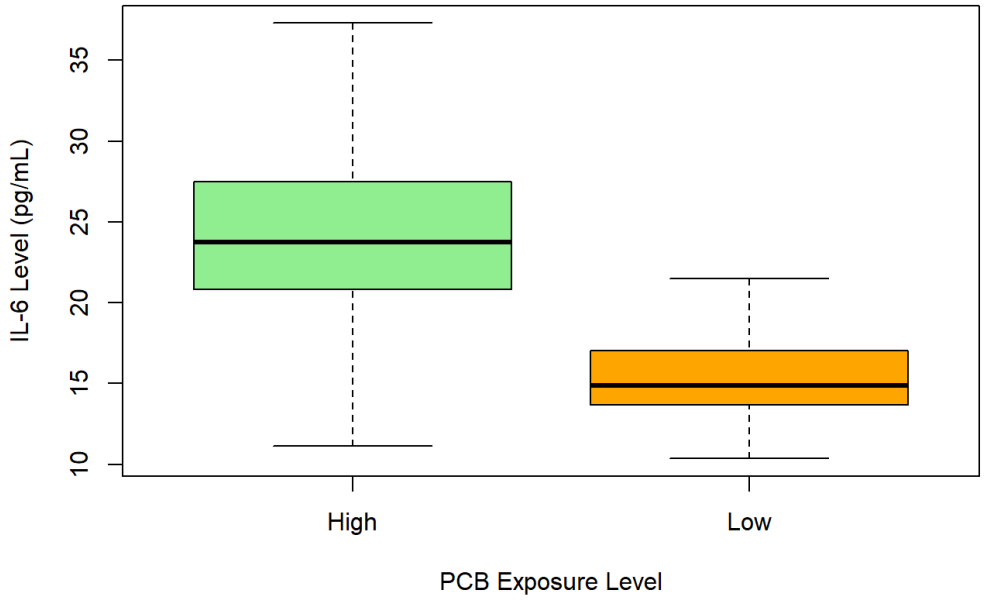
- Grafičke metode (boxplot)



- Statistički testovi (Levene-ov test, Bartlettov test)

Testiranje srednje vrijednosti razine citokina (IL-6)

- Provjera pretpostavki o homogenosti varijance:




```
### Levene's test
```{r}

leveneTest(CytokineLevel ~ Exposure, data = cytokine_data)
```
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value   Pr(>F)
group  1  9.0952 0.004551 **
      38
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Studentov t-test za nejednake varijance (Welchov test)

$$t_{df} = \frac{\bar{X}_1 - \bar{X}_2}{SE_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

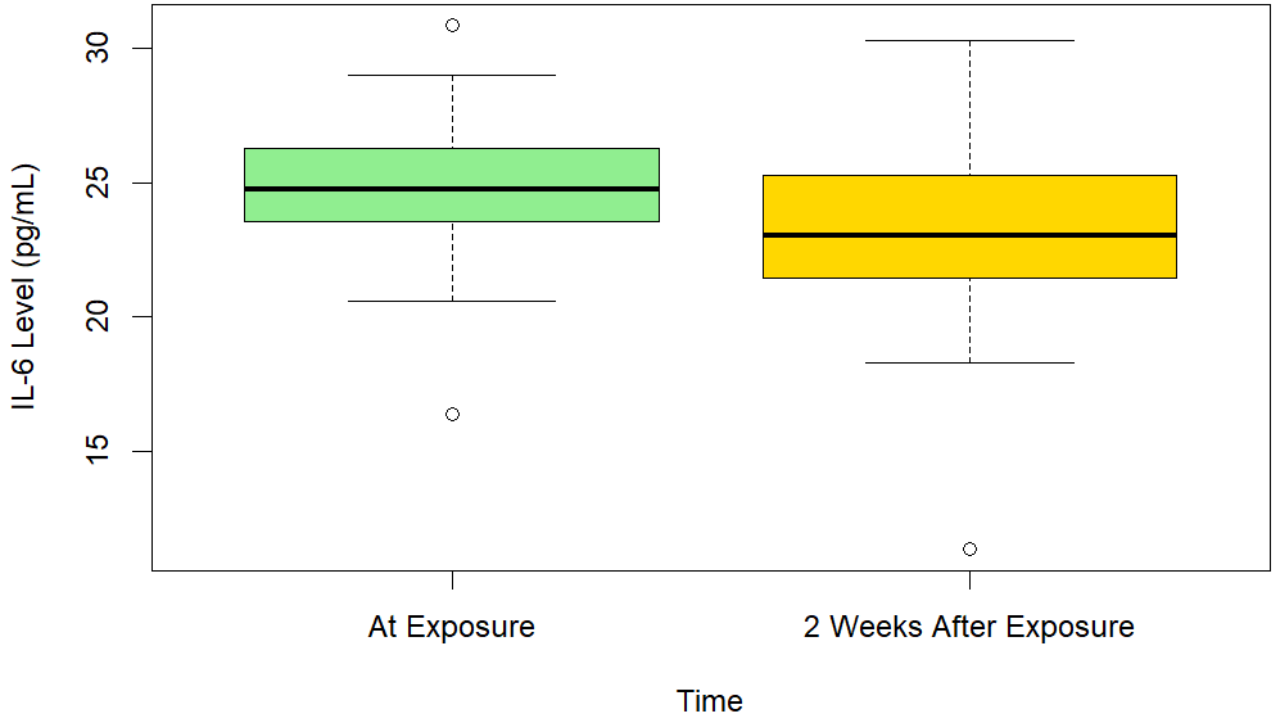
```
# T-test for cytokine levels
t_test_cytokine <- t.test(CytokineLevel ~ Exposure, data = cytokine_data, var.equal =
F)
print(t_test_cytokine)
```

```
##
##  Welch Two Sample t-test
##
## data:  CytokineLevel by Exposure
## t = 6.159, df = 27.686, p-value = 1.253e-06
## alternative hypothesis: true difference in means between group High and group Low is
not equal to 0
## 95 percent confidence interval:
##   5.979168 11.942916
## sample estimates:
## mean in group High  mean in group Low
##           24.28050           15.31946
```

Odbacujemo nultu hipotezu na razini značajnosti 0.05!

Što ako uzorci nisu međusobno neovisni

- Npr. Mjerimo promjenu u razini citokina u vrijeme izlaganja PCB-u i 2 tjedna nakon izlaganja PCB-u
- Hipoteza: Postoji razlika u razini IL-6 ovisno o vremenu mjerenja



Studentov t-test za povezane uzorke

- Kada uspoređujemo dva povezana uzorka često nas ne zanimaju apsolutne vrijednosti već razlika između izmjerenih vrijednosti za svaki par observacija u povezanih uzorcima

Uzorak 1	Uzorak 2	Razlika
X_{11}	X_{21}	$X_{11} - X_{21}$
X_{12}	X_{22}	$X_{12} - X_{22}$
.	.	.
.	.	.
.	.	.
X_{1n}	X_{2n}	$X_{1n} - X_{2n}$

- Uzorak razlika u mjerenjima uzet iz populacije razlika
- Karakterizira ga srednja vrijednost i standardna devijacija
- Srednja vrijednost razlika u mjerenjima $\bar{D}$

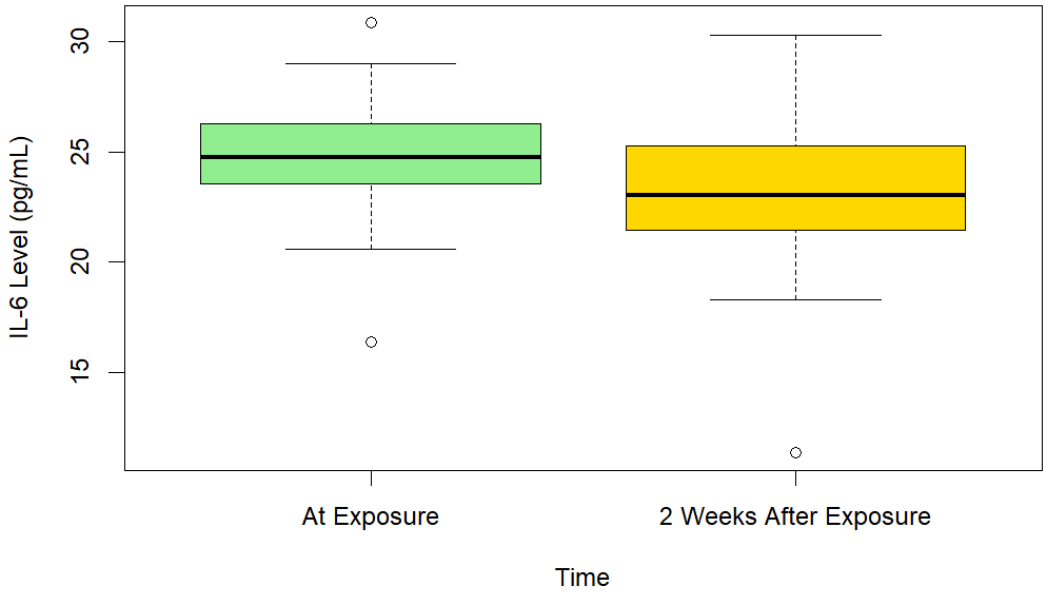
Studentov t-test za povezane uzorke

- Nulta hipoteza: Nema razlike među dva međusobno ovisna mjerenja u našem uzorku
- Zbog varijabilnosti očekujemo da će neke razlike biti pozitivne, neke negativne, ali srednja vrijednost tih razlika biti će jednaka nuli

$$t_{n-1} = \frac{\bar{D} - \mu_{\bar{D}}}{SE_{\bar{D}}}$$

- Koristimo t-test kako bismo utvrdili je li srednja vrijednost razlika značajno različita od nule
- Naš uzorak je zapravo stupac s razlikama u mjerenjima na kojem ćemo napraviti t-test na jednom uzorku s nultom hipotezom da je $\mu = 0$

Studentov t-test za povezane uzorke




```
##  Welch Two Sample t-test
##
## data:  CytokineLevel by Exposure
## t = 1.3661, df = 35.587, p-value = 0.1805
## alternative hypothesis: true difference in means betwee
n group At Exposure and group 2 Weeks After Exposure is no
t equal to 0
## 95 percent confidence interval:
##  -0.7668215  3.9280175
## sample estimates:
##            mean in group At Exposure mean in group 2 We
eks After Exposure
##                              24.7520
23.1714
```

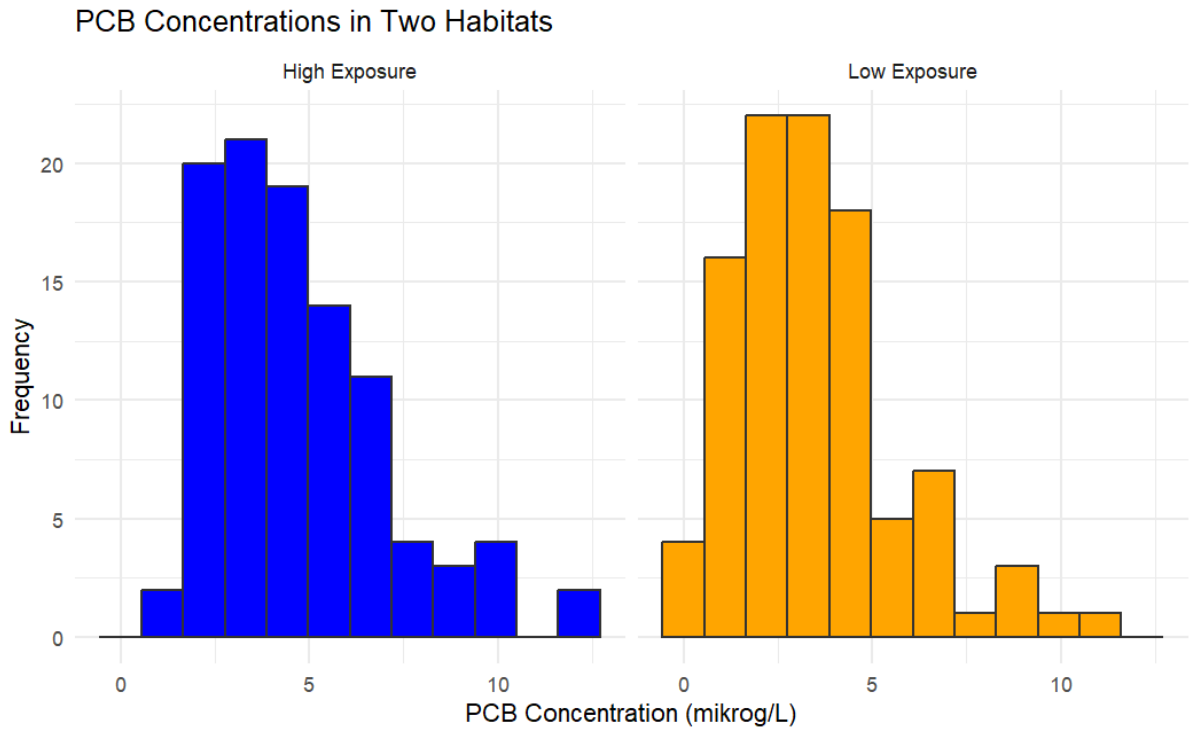
Ne odbacujemo nultu hipotezu na razini značajnosti 0.05!

```
##  Paired t-test
##
## data:  CytokineLevel by Exposure
## t = 3.374, df = 19, p-value = 0.003186
## alternative hypothesis: true mean difference is not equ
al to 0
## 95 percent confidence interval:
##  0.6001007 2.5610954
## sample estimates:
## mean difference
##        1.580598
```

Odbacujemo nultu hipotezu na razini značajnosti 0.05!

Što kad podaci nisu normalno distribuirani?

- Želimo testirati koncentraciju PCB-a u staništima koja jesu/nisu izložena zagađenju.
- Koncentracija PCB-a nije normalno distribuirana.



Neparametarski testovi

- Testovi koji nemaju pretpostavki o distribuciji proučavanih varijabli
- Testiraju se rangovi podataka, a ne same vrijednosti podataka
- Mogu se koristiti za male uzorke (N<30) i za varijable koje ne slijede normalnu distribuciju (!)
- Manja statistička snaga testa

- **Wilcoxon-Mann-Whitney rank-sum test (Mann-Whitney U)** – usporedba dva neovisna uzorka
- **Wilcoxon signed rank test** – usporedba dva ovisna uzorka

Wilcoxon-Mann-Whitney Rank Sum Test

- Test statistika računa se na rangovima, a ne sirovim podacima

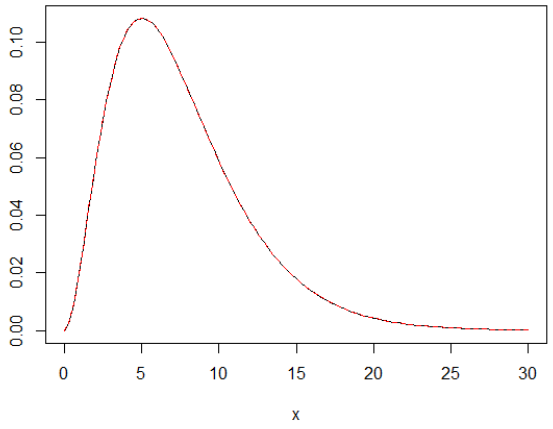
- Hipoteza
  - H_0 – dvije populacije imaju identičnu distribuciju
  - H_1 – “lokacije populacija se razlikuju”

- Pretpostavke:
  - Podaci dolaze iz nasumičnih uzoraka
  - Podaci unutar uzoraka su međusobno neovisni
  - Uzorci su međusobno neovisni

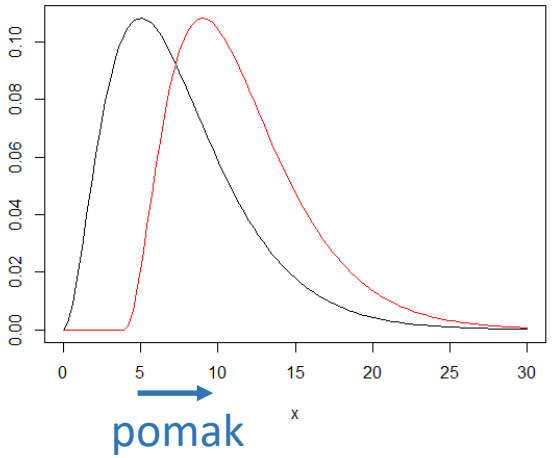
Wilcoxon-Mann-Whitney Rank Sum Test

- Wilcoxon -Mann-Whitney Rank Sum Test pokušava otkriti pomake u lokaciji
  - H1 : A > B (A pomaknuta u desno od B)
  - H1 : A < B (A pomaknuta u lijevo od B)
  - H1 : A ≠B

H_0: A = B

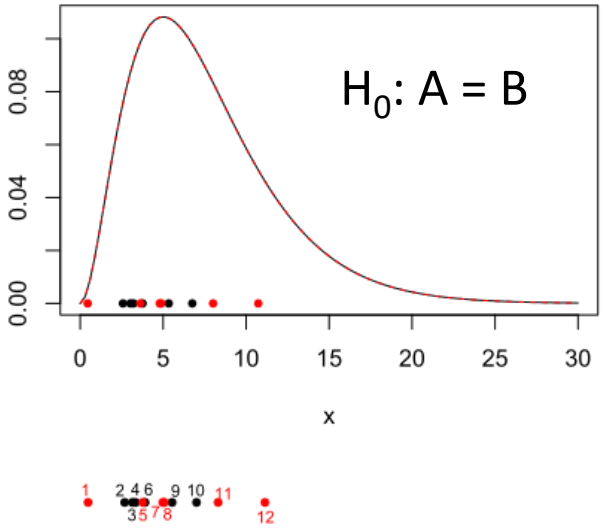


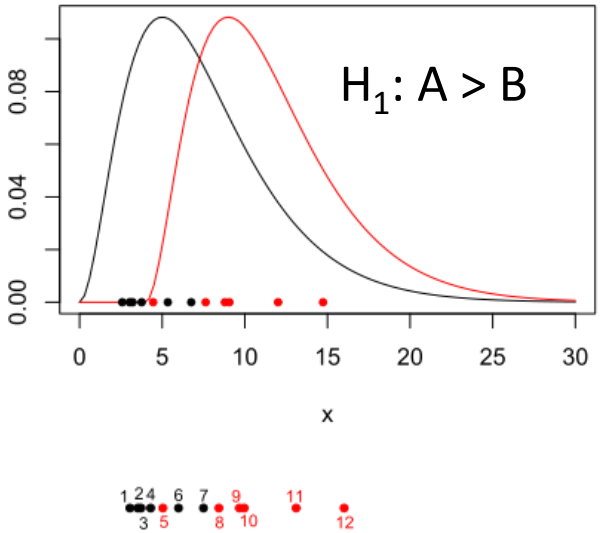
H_1: A > B



Wilcoxon-Mann-Whitney Rank Sum Test

- Staviti sve observacije u jedan uzorak i rangirati $n_A + n_B$ observacija kombiniranog uzorka
- Wilcoxon rank-sum statistika – suma rangova observacija iz jednog od uzoraka
- w_A = suma rangova observacija iz uzorka A
- a) H_0:A=B b) H_1: A>B

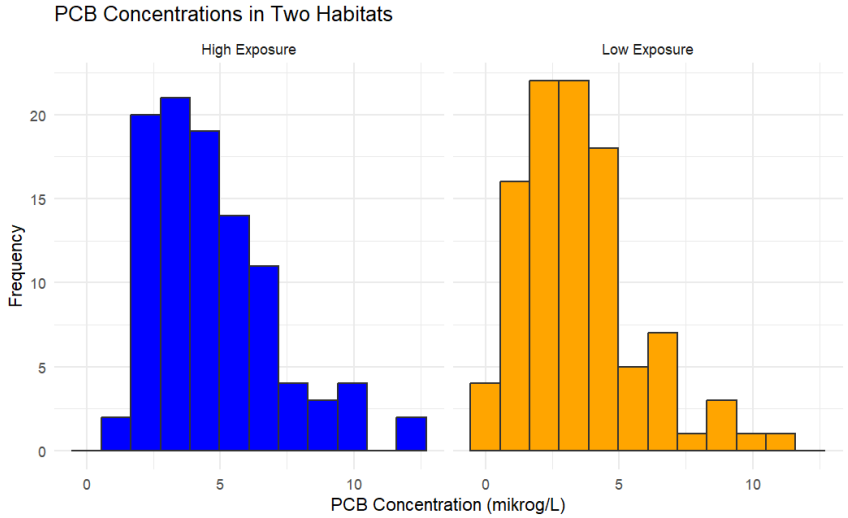




$$(H_1 : A > B) \qquad P\text{-}value = \Pr(W_A \geq w_A)$$

$$(H_1 : A < B) \qquad P\text{-}value = \Pr(W_A \leq w_A)$$

Wilcoxon-Mann-Whitney Rank Sum Test




```
# Test difference in distributions using a non-parametric test (Wilc
oxon-Mann-Whitney U Test)
wilcox_test_result <- wilcox.test(Concentration ~ Habitat, data = pc
b_data)
print(wilcox_test_result)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Concentration by Habitat
## W = 6697, p-value = 3.395e-05
## alternative hypothesis: true location shift is not equal to 0
```

```
# Test difference in means using a t-test
t_test_result <- t.test(Concentration ~ Habitat, data = pcb_data)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  Concentration by Habitat
## t = 3.9501, df = 197.54, p-value = 0.0001086
## alternative hypothesis: true difference in means between group Hi
gh Exposure and group Low Exposure is not equal to 0
## 95 percent confidence interval:
##  0.6405625 1.9177830
## sample estimates:
## mean in group High Exposure  mean in group Low Exposure
##                    4.722172                    3.442999
```

Wilcoxon signed rank test

- Verzija prethodnog testa gdje se pojedinci mogu mjeriti dvaput ili se mogu uzeti u obzir mjerenja prije i poslije (spareni test)
- Podaci koji imaju isti broj mjerenja.
- Upareni uzorci $(X_i,Y_i) \rightarrow D_i=X_i-Y_i$
- Ako niti jedan tretman nema učinka, tada ne samo da bi razlike trebale biti jednako raspoređene s obje strane 0, nego bi i to koliko su razlike udaljene od 0 trebale biti isto s obje strane.
- Postupak:
  - Izračunajte D_i i rangirati
  - Izračunajte W+ = zbroj rangova s pozitivnim predznakom ili W− = zbroj rangova s negativnim predznakom
  - Idejaako je distribucija X (F(x)) ista kao distribucija Y (F(y)), tada je jednako vjerojatno da će Di biti pozitivni kao i negativni. Dakle, oko pola rangova je pozitivno, a pola je negativno.ako je F(y) veći od F(x), očekujte da većina rangova ima pozitivne predznake i stoga će W+ biti velik
- Za $n \geq 20$ približno normalno raspoređen

Kada koristiti neparametarske testove

- Podaci koji nisu normalno distribuirani (jako asimetrične distribucije)
- Podaci koji sadrže stršeće vrijednosti (*outliere*)
- Mali broj podataka

Tipovi varijabli:

- Numeričke (kvantitativne) - diskretne i kontinuirane
- Kategoričke (kvalitativne) - nominalne i ordinalne

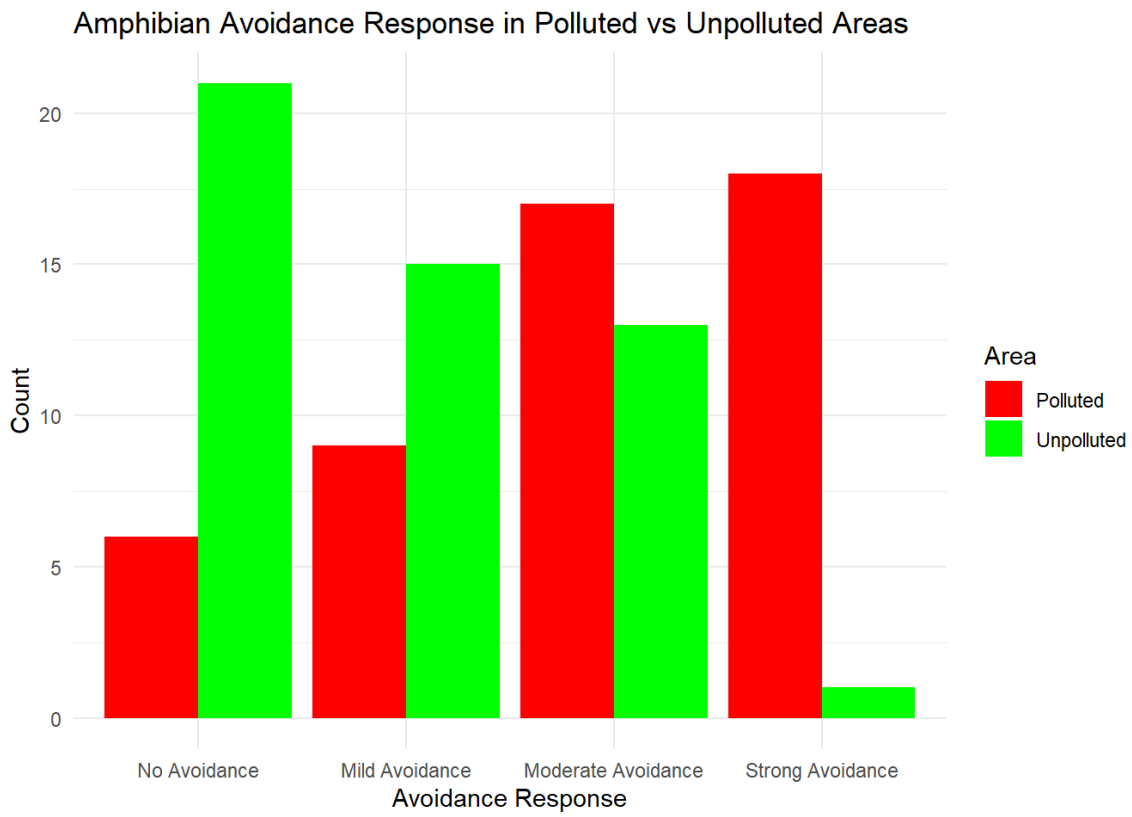
Neparametarski testovi za ordinalne variable

- Izbjegavanje kontaminiranog područja

1. Bez izbjegavanja: Vodozemci ostaju u kontaminiranoj zoni unatoč prisutnosti PCB-a.
2. Blago izbjegavanje: Vodozemci pokazuju povremeno kretanje izvan kontaminirane zone.
3. Umjereno izbjegavanje: Vodozemci često izbjegavaju kontaminirano područje, ali se ipak povremeno vraćaju.
4. Strogo izbjegavanje: Vodozemci rijetko ili nikad ne ulaze u kontaminirano područje.

	Polluted	Unpolluted
No Avoidance	6	21
Mild Avoidance	9	15
Moderate Avoidance	17	13
Strong Avoidance	18	1

Neparametarski testovi za ordinalne variable



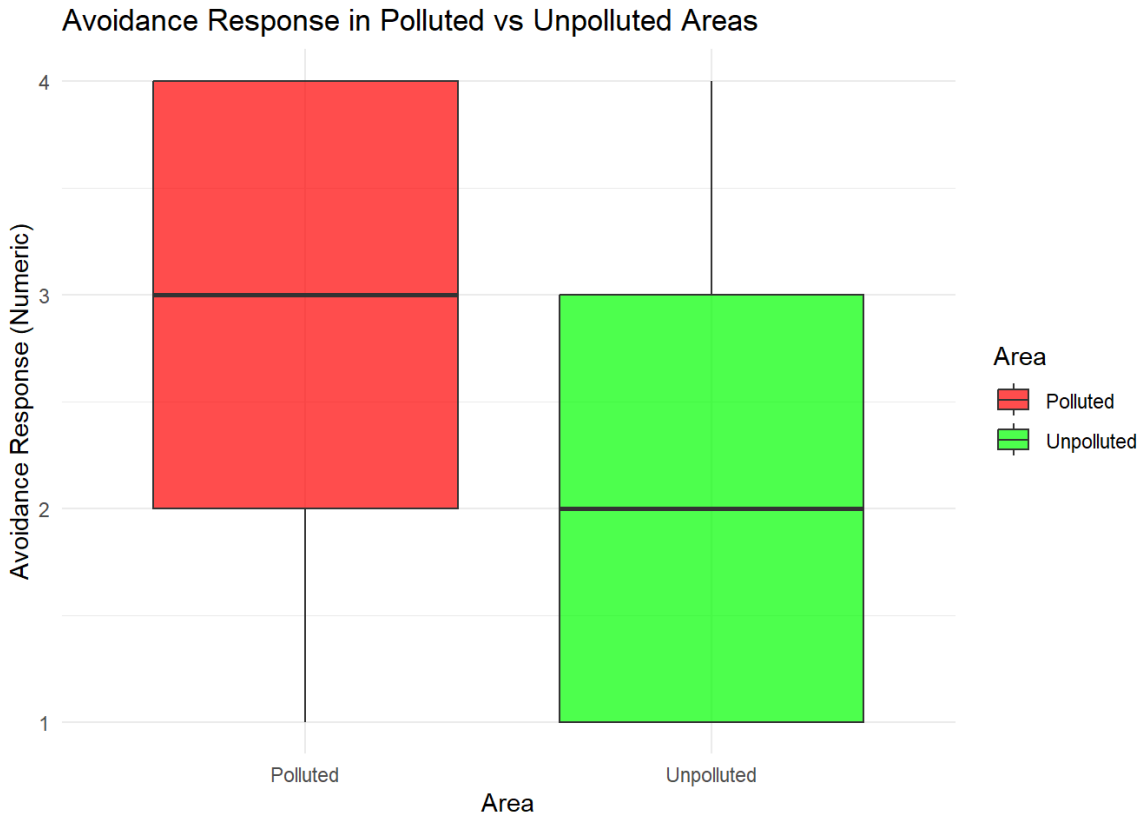
- Vrijednosti moramo pretvoriti u numeričke
- Udaljenosti između brojeva 1, 2, 3, 4 kod ordinalne varijable ne moraju biti jednake kao kod numeričke varijable

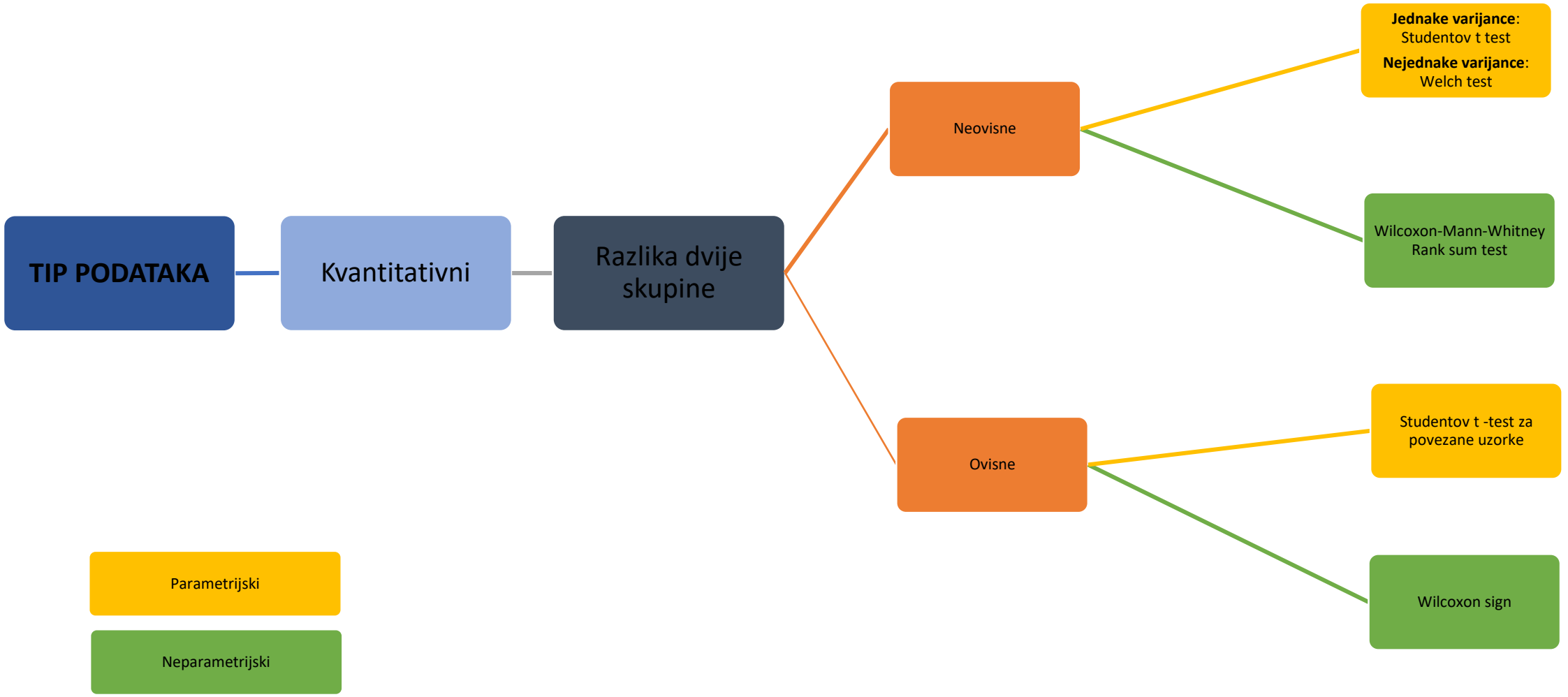
Neparametarski testovi za ordinalne variable

```
# Convert ordinal variable to numeric for Wilcoxon test
amphibian_data$Avoidance_Numeric <- as.numeric(amphibian_data$Avoidance)

# Wilcoxon rank-sum test
test_result <- wilcox.test(Avoidance_Numeric ~ Area, data = amphibian_data)
print(test_result)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Avoidance_Numeric by Area
## W = 1933, p-value = 1.106e-06
## alternative hypothesis: true location shift is not equal to 0
```




TIP PODATAKA
Kvantitativni
Razlika dvije skupine
Neovisne
Jednake varijance: Studentov t test
Nejednake varijance: Welch test
Wilcoxon-Mann-Whitney Rank sum test
Ovisne
Studentov t -test za povezane uzorke
Wilcoxon sign
Parametrijski
Neparametrijski

Ostale neparametarske metode

- Kruskal-Wallis test, Friedmanov test – neparametarska ANOVA
- Spearmanov rho i Kendallov tau – naparametarska korelacija
- Ordinalna regresija – neparametarska regresija