Laboratorijska vježba 3 - Statističko zaključivanje dvije varijable

Metode statističke analize podataka

2025./2026.

Cilj vježbe:

- Uvježbati primjenu parametarskih i neparametarskih metoda za usporedbu dvaju uzoraka (zavisnih i nezavisnih).
- Na temelju provjera pretpostavki (normalnost, homoskedastičnost) odabrati odgovarajući statistički test.
- Primijeniti metode jednostavne linearne regresije za modeliranje odnosa između dviju varijabli.
- Provesti analizu reziduala i interpretirati koeficijent determinacije (R^2) kao mjeru adekvatnosti modela.

Napomena za studente: Za uspješno polaganje **LV3** potrebno je:

- proći gradivo s predavanja koje obuhvaća statističko zaključivanje o dvjema varijablama i osnovne koncepte linearne regresije,
- proučiti **predložak laboratorijske vježbe** i razumjeti teorijski uvod (pretpostavke testova, tumačenje p-vrijednosti i R^2),
- pregledati i proći **Python notebook s predavanja** (primjeri testova i regresije).

1 Uvod

Statističko zaključivanje dviju varijabli (engl. *bivariate statistical inference*) bavi se procjenom odnosa ili razlika između dviju populacija/varijabli na temelju uzorka:

1. usporedba dviju grupa (razlika očekivanja)
2. modeliranje odnosa (korelacija/regresija).

Ključni su pojmovi uzorkovanje, pogreška uzorkovanja, intervali pouzdanosti i testiranje hipoteza.

Pretpostavka homoskedastičnosti: Homoskedastičnost znači da su varijance uspoređivanih skupina približno jednake. Ako ta pretpostavka nije zadovoljena, koristi se test koji ne pretpostavlja jednakost varijanci — primjerice Welchov t-test.
Ključne metode:

1. usporedba srednjih vrijednosti grupa (npr. t-testovi, Z-testovi, neparametrijske metode).
2. modeliranje odnosa među varijablama (korelacija i jednostavna linearna regresija).

1.1 Usporedba dviju grupa

Cilj je utvrditi postoji li statistički značajna razlika između srednjih vrijednosti dviju skupina. Prije odabira testa ključno je analizirati:

- Nezavisnost ili zavisnost uzoraka
- Normalnost podataka
- Homoskedastičnost varijanci

1.1.1 Parametarski testovi (uz pretpostavku normalnosti)

Parametarski testovi zahtijevaju da podaci dolaze iz normalno distribuiranih populacija, ili da su uzorci dovoljno veliki, barem $n \geq 30$ ali bolje $n \geq 40$, za primjenu centralnog graničnog teorema.

Table 1: Vrste t-testova i njihove pretpostavke

Tip uzorka	**Odabrani test**	**Pretpostavke**
Nezavisni uzorci	t-test za nezavisne uzorke	Normalna distribucija; homoskedastičnost (jednake varijance, što dovodi do *Pooled t-testa*); ako varijance nisu jednake, koristi se *Welchov t-test* ($\sigma_1^2 \neq \sigma_2^2$).
Zavisni uzorci	Parni t-test (*Paired t-test*)	Analiza se fokusira na razlike ($D_j = X_{1j} - X_{2j}$). Pretpostavka je normalna distribucija tih razlika.

1.1.2 Razlika između Pooled i Welchovog t-testa

Razlika između t-testa s parametrom `equal_var=True` (Pooled t-test) i t-testa s `equal_var=False` (Welchov t-test) je u pretpostavkama i načinu računanja statistike:

Pooled t-test (equal_var=True): Pretpostavlja da obje grupe imaju jednake varijance (homoskedastičnost). Varijance se na ovaj način "povlače" ili objedinjuju (*pooled variance*) u jednu zajedničku procjenu varijance koja se koristi u testu. Ovaj test je optimalan ako je pretpostavka o jednakosti varijanci zadovoljena, ali nije robustan ako varijance nisu jednake.

Welchov t-test (equal_var=False): Ne pretpostavlja jednaku varijancu među grupama (heteroskedastičnost dopuštena). Koristi različite procjene varijanci za svaku grupu i računa modificirane stupnjeve slobode pomoću Welch-Satterthwaite aproksimacije. Ovaj test je robusniji i sigurniji kada varijance nisu jednake, ali može imati nešto manju snagu ako su varijance zaista jednake.

U praksi, ako Leveneov ili sličan test pokaže da su varijance jednake, preporučuje se Pooled t-test (`equal_var=True`), a ako nisu, sigurnije je koristiti Welchov t-test (`equal_var=False`).

Table 2: Usporedba Pooled i Welchovog t-testa

Test	**Pretpostavka o varijancama**	**Varijanca obračunata kao**	**Stupnjevi slobode**	**Kada koristiti**
Pooled t-test	Jednake varijance	Zajednička (*pooled*) varijanca	$v = n_1 + n_2 - 2$	Kad se može pretpostaviti da su varijance jednake (homoskedastičnost).
Welchov t-test	Nejednake varijance	Varijance procijenjene odvojeno za svaku skupinu	Aproksimirani (WS formula)	Kad varijance nisu jednake ili je homoskedastičnost upitna.

Ovakav pristup omogućava preciznije zaključke o razlikama između grupa ovisno o zadovoljenju pretpostavki. **Python primjer — t-test za nezavisne uzorke:**

```python
import numpy as np
from scipy import stats

# Podaci o prinosu dvaju katalizatora
kat1 = np.array([91.5, 94.2, 92.2, 95.4, 91.8, 89.1, 94.7, 89.2])
kat2 = np.array([89.2, 91.0, 90.5, 93.2, 97.2, 97.0, 91.1, 92.8])

# Provjera jednakosti varijanci (Leveneov test)
levene = stats.levene(kat1, kat2)
print(f"Levene p-vrijednost = {levene.pvalue:.3f}")

# Odabir testa prema rezultatu
equal_var = levene.pvalue > 0.05
t, p = stats.ttest_ind(kat1, kat2, equal_var=equal_var)

print(f"T-statistika = {t:.2f}, p-vrijednost = {p:.3f}")
if p < 0.05:
print("Zakljucak: Postoji znacajna razlika izmedju skupina.")
else:
print("Zakljucak: Nema znacajne razlike izmedju skupina.")
```

1.1.3 Neparametarski testovi (kada normalnost nije zadovoljena)

Kada pretpostavka normalnosti nije zadovoljena, koriste se neparametrijske alternative. Ovi testovi mjere razliku u lokaciji (medijanima) distribucija, a ne strogo u srednjim vrijednostima.

Table 3: Neparametrijski testovi i njihove analogije s t-testovima

Tip uzorka	**Odabrani test**	**Analogija s t-testom**
Nezavisni uzorci	Mann–Whitney U test (ili Wilcoxon rank-sum test)	t-test za nezavisne uzorke
Zavisni uzorci	Wilcoxon signed-rank test	t-test za zavisne uzorke

Primjer: Wilcoxonov test za zavisne uzorke

Wilcoxonov test koristi se kao **neparametarska alternativa** *parnom t-testu* kada **razlike između parova nisu normalno distribuirane**. Mjeri razlikuju li se medijani dviju povezanih skupina (npr. prije i poslije intervencije).

Hipoteze:

$$H_0 : \text{medijan razlika} = 0 \quad \text{(nema promjene)}$$
$$H_1 : \text{medijan razlika} \neq 0 \quad \text{(postoji promjena)}$$

Primjer situacije: Deset zaposlenika ispunilo je upitnik o stresu prije i nakon kratkog programa opuštanja. Želimo utvrditi postoji li statistički značajna promjena u razini stresa.

```
import numpy as np
from scipy.stats import wilcoxon, shapiro

# Rezultati prije i poslije programa (manji broj = manji stres)
prije = np.array([22, 25, 20, 18, 24, 19, 23, 21, 26, 20])
poslije= np.array([18, 20, 19, 16, 21, 17, 20, 19, 22, 18])

# 1) Izracun razlika
D = prije - poslije
print("Razlike:", D)

# 2) Provjera normalnosti razlika (Shapiro-Wilk)
sh = shapiro(D)
print(f"Shapiro p-vrijednost = {sh.pvalue:.3f}")

# 3) Ako nije normalno (p < 0.05), koristimo Wilcoxonov test
res = wilcoxon(prije, poslije)
print(f"Wilcoxon statistika = {res.statistic}, p-vrijednost = {res.pvalue:.4f}")

# 4) Zakljucak
```

```
    alpha = 0.05
    if res.pvalue < alpha:
    print("Zakljucak: Odbacujemo H0 -> postoji znacajna razlika u razini stresa.")
    else:
    print("Zakljucak: Ne odbacujemo H0 -> nema dokaza o promjeni u razini stresa.")
```

Tumačenje: Ako je p-vrijednost manja od 0.05, možemo zaključiti da postoji statistički značajna promjena u rezultatima prije i poslije intervencije. Wilcoxonov test ne zahtijeva normalnost podataka, ali pretpostavlja da su razlike simetrične oko medijana.

Napomena: Ovaj test prikladan je i za male uzorke, ali za veće skupove s vezanim rangovima SciPy koristi normalnu aproksimaciju. Kod interpretacije naglasiti da test provjerava *promjenu u medijanu*, a ne u srednjoj vrijednosti.

1.2 Jednostavna linearna regresija i korelacija

Regresijska analiza je statistički alat za istraživanje i modeliranje odnosa između varijabli koje su povezane na nedeterministički način.

1.2.1 Model i procjena

Model jednostavne linearne regresije koristi jednu nezavisnu varijablu x i zavisnu varijablu Y

Model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Koeficijenti procjenjuju metodom najmanjih kvadrata (Least Squares).
- Reziduali: $e_i = y_i - \hat{y}_i$.
- Test značajnosti nagiba: $H_0 : \beta_1 = 0$ - nema linarnog odnos
- Koeficijent determinacije R^2: udio varijance objašnjen modelom.
- Intervali: predviđanja (PI) i pouzdanosti (CI).

1.2.2 Analiza reziduala i provjera homoskedastičnosti

U modelu linearne regresije

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

pretpostavlja se da pogreške (reziduali) imaju konstantnu varijancu za sve vrijednosti x, tj. da vrijedi

$$Var(\varepsilon_i) = \sigma^2 \quad \text{za sve } i.$$

Ta pretpostavka naziva se **homoskedastičnost**. Ako varijanca pogrešaka raste ili pada s x, govorimo o **heteroskedastičnosti**.

1. Vizualna provjera (reziduali vs. predviđene vrijednosti) Ako su reziduali slučajno raspoređeni oko nule, slične raspršenosti za sve predviđene vrijednosti, pretpostavka je zadovoljena. Pojava „lijevka" (povećanje rasapa s rastom x) ukazuje na heteroskedastičnost.

```
# Pretpostavimo da su izracunati predvidjeni y_pred i reziduali
plt.scatter(y_pred, residuals)
plt.axhline(0, color='red', linestyle='--')
plt.xlabel("Predvidjene vrijednosti")
plt.ylabel("Reziduali")
plt.title("Provjera homoskedasticnosti (vizualno)")
plt.show()
```

2. Numerička provjera (Leveneov test) Leveneov test ispituje jednakost varijanci reziduala u različitim područjima predviđenih vrijednosti. Ako je $p > 0.05$, pretpostavka homoskedastičnosti nije narušena.

```
import pandas as pd
from scipy.stats import levene

# Podjela reziduala u dvije grupe prema medijanu predvidjenih vrijednosti
group = pd.qcut(y_pred, 2, labels=False)
res1 = residuals[group == 0]
res2 = residuals[group == 1]

stat, p = levene(res1, res2)
print(f"Levene p-vrijednost = {p:.3f}")
if p > 0.05:
print("Zakljucak: Varijance su jednake (homoskedasticnost zadovoljena).")
else:
print("Zakljucak: Varijance se razlikuju (heteroskedasticnost).")
```

3. Zaključak Homoskedastičnost je važna jer utječe na pouzdanost procjene standardnih pogrešaka i t-testova za koeficijente. Ako nije zadovoljena, može se:

- primijeniti transformacija zavisne varijable (npr. $\log(Y)$, $\sqrt{Y}$),
- koristiti robusne standardne pogreške (*robust standard errors*),
- ili preispitati prikladnost linearnog modela.

2 Riješeni primjeri

ZADATAK 1: Nezavisni uzorci

Uspoređuje se **vrijeme izvršavanja (ms)** dvaju algoritama na istom skupu zadataka. Pretpostavite da su pokusi nezavisni po algoritmu.

Podaci (primjer) *Algoritam A:*
`[52.1, 48.9, 51.7, 50.8, 49.5, 52.9, 47.8, 50.3, 49.9, 51.2, 48.4, 50.1, 49.6, 52.0, 50.6]`
Algoritam B:
`[55.4, 54.1, 56.2, 53.9, 55.0, 57.1, 54.6, 56.4, 55.8, 54.9, 56.0, 55.2, 56.3, 54.7, 55.6]`

Zadatak.

1. Izračunajte sredine i SD. Grafički prikažite podatke i objasnite dobiveni prikaz.
2. Testirajte **homoskedastičnost** (Levene). Ako $p < 0{,}05 \Rightarrow$ heteroskedastičnost.
3. Postavite $H_0 : \mu_A = \mu_B$ (dvosmjerno). Odaberite *pooled* ili *Welch* t-test prema Levene.
4. Provedite test za $\alpha = 0{,}05$ i $\alpha = 0{,}01$. Komentirajte razlike zaključka.
5. Izračunajte 95% CI za $\mu_A - \mu_B$. Objasnite odnos CI i testa.

```python
import numpy as np
from scipy import stats

A = np.array([52.1,48.9,51.7,50.8,49.5,52.9,47.8,
50.3,49.9,51.2,48.4,50.1,49.6,52.0,50.6])
B = np.array([55.4,54.1,56.2,53.9,55.0,57.1,54.6,
56.4,55.8,54.9,56.0,55.2,56.3,54.7,55.6])

# 1) Deskriptivna statistika (dio samo, ostalo rijesiti iz LV1)
print(np.mean(A), np.std(A, ddof=1))
print(np.mean(B), np.std(B, ddof=1))

# 2) Homoskedasticnost (Levene): H0 = jednake varijance
lev = stats.levene(A, B)
equal_var = lev.pvalue >= 0.05
print("Levene p =", round(lev.pvalue,3), " -> equal_var =", equal_var)

# 3) t-test (pooled ako equal_var=True, inace Welch)
t,p = stats.ttest_ind(A, B, equal_var=equal_var)
print("t =", round(t,2), " p =", round(p,5))

# 4) Zakljucci za dvije razine
for alpha in (0.05, 0.01):
print(f"alpha={alpha}: ", "Odbacujemo H0" if p<alpha else "NE odbacujemo H0")

# 5) 95% CI za (mu_A - mu_B)
n1, n2 = len(A), len(B)
s1, s2 = np.var(A, ddof=1), np.var(B, ddof=1)
```

```
if equal_var:
sp2 = ((n1-1)*s1 + (n2-1)*s2) / (n1+n2-2)
se = np.sqrt(sp2*(1/n1 + 1/n2))
df = n1+n2-2
else:
se = np.sqrt(s1/n1 + s2/n2)
df_num = (s1/n1 + s2/n2)**2
df_den = (s1**2/((n1**2)*(n1-1))) + (s2**2/((n2**2)*(n2-1)))
df = df_num/df_den # Welch df

mean_diff = np.mean(A) - np.mean(B)
tcrit = stats.t.ppf(1-0.05/2, df)
ci = (mean_diff - tcrit*se, mean_diff + tcrit*se)
print("95\% CI (mu_A - mu_B) =", tuple(round(x,3) for x in ci)
    )
```

Mann–Whitney U (ako podaci nisu normalno distribuirani):

```
res = stats.mannwhitneyu(A, B, alternative="two-sided", method
    ="auto")
print("U =", res.statistic, " p =", res.pvalue)
```

ZADATAK 2: Zavisni (parni) uzorci (prije $\to$ poslije)

Ispitanici rješavaju test **prije** i **poslije** kratke radionice. Želimo utvrditi postoji li promjena.

Podaci (primjer) `Prije: [47, 50, 52, 41, 44, 53, 48, 46, 55, 49, 51, 43]`
`Poslije: [52, 54, 55, 45, 48, 57, 51, 49, 58, 52, 53, 47]`

Zadatak.

1. Izračunajte razlike $D =$ Prije $-$ Poslije; opišite ih (sredina, SD).
2. Provjerite normalnost *razlika* (Shapiro–Wilk).
3. Ako su razlike normalne $\Rightarrow$ *parni t-test*; inače *Wilcoxon signed-rank*.
4. Napišite **Zaključak** (dvosmjerno, $\alpha = 0{,}05$).

```
import numpy as np
from scipy import stats

prije = np.array([47,50,52,41,44,53,48,46,55,49,51,43])
poslije= np.array([52,54,55,45,48,57,51,49,58,52,53,47])

D = prije - poslije
print("Mean(D) =", np.mean(D), " SD(D) =", np.std(D, ddof=1))

# Normalnost razlika
```

```
    sh = stats.shapiro(D)
    print("Shapiro p =", sh.pvalue)

    alpha = 0.05
    if sh.pvalue >= alpha:
    t,p = stats.ttest_rel(prije, poslije)
    print("Paired t-test: t =", round(t,2), " p =", round(p,5))
    print("Zakljucak:", "Odbacujemo H0" if p<alpha else "NE
       odbacujemo H0")
    else:
    res = stats.wilcoxon(prije, poslije)
    print("Wilcoxon: stat =", res.statistic, " p =", round(res.
       pvalue,5))
    print("Zakljucak:", "Odbacujemo H0" if res.pvalue<alpha else "
       NE odbacujemo H0")
```

ZADATAK 3: Jednostavna linearna regresija i analiza reziduala

Modeliramo odnos između **čistoće kisika** (Y) i **postotka ugljikovodika** (x) u zraku, na temelju podataka iz **Tablice 11.1** (Montgomery et al.). Cilj je procijeniti jednostavni linearni model:

$$y = \beta_0 + \beta_1 x$$

te ispitati postoji li značajna linearna povezanost između varijabli.

Zadatak:

1. Procijenite koeficijente $\hat{\beta}_0$ i $\hat{\beta}_1$ metodom najmanjih kvadrata koristeći funkciju `linregress`.
2. Ispišite jednadžbu regresijskog pravca, koeficijent korelacije r, determinacije R^2 te p-vrijednost za test $H_0 : \beta_1 = 0$.
3. Prikažite scatter dijagram podataka s ucrtanim regresijskim pravcem.
4. Analizirajte reziduale i provjerite pretpostavku homoskedastičnosti (reziduali nasumično raspršeni oko nule).
5. Predvidite čistoću kisika za $x_0 = 1{,}00$ % te interpretirajte dobiveni rezultat.

Podaci (20 opažanja, Tablica 11.1) `x = [0.99, 1.02, 1.15, 1.29, 1.46, 1.36, 0.87, 1.23, 1.55, 1.40,`

```
    1.19, 1.15, 0.98, 1.01, 1.11, 1.20, 1.26, 1.32, 1.43, 0.95]
y = [90.01, 89.05, 91.43, 93.74, 96.73, 94.45, 87.59, 91.77, 99.42,
93.65,
    93.54, 92.52, 90.56, 89.54, 89.85, 90.39, 93.25, 93.41, 94.98,
87.33]
```

y

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import linregress

# Podaci (primjer iz tablice 11.1)
x = np.array([0.99, 1.02, 1.15, 1.29, 1.46, 1.36, 0.87, 1.23, 1.55, 1.40])
y = np.array([90.01, 89.05, 91.43, 93.74, 96.73, 94.45, 87.59, 91.77, 99.42, 93.65])

# 1) Procjena modela
slope, intercept, r, p, se = linregress(x, y)

print(f"Jednadzba: y = {intercept:.3f} + {slope:.3f}x")
print(f"Koeficijent korelacije r = {r:.3f}")
print(f"Koeficijent determinacije R^2 = {r**2:.3f}")
print(f"P-vrijednost za nagib = {p:.4f}")

# 2) Graficki prikaz
plt.scatter(x, y, label="Opazeni podaci")
plt.plot(x, intercept + slope*x, color="red", label="Regresijski pravac")
plt.xlabel("Ugljikovodici (%)")
plt.ylabel("Cistoca kisika (%)")
plt.legend()
plt.show()

# 3) Analiza reziduala
y_pred = intercept + slope*x
residuals = y - y_pred

plt.scatter(y_pred, residuals)
plt.axhline(0, color="red", linestyle="--")
plt.xlabel("Predvidjene vrijednosti")
plt.ylabel("Reziduali")
plt.title("Provjera homoskedasticnosti")
plt.show()

# 4) Predikcija za novu vrijednost x0
x0 = 1.00
y0 = intercept + slope*x0
print(f"Predvidjena cistoca za x0={x0:.2f}%: {y0:.3f}%")

# Zakljucak
if p < 0.05:
print("Zakljucak: Odbacujemo H0. Postoji znacajna linearna povezanost izmedju varijabli.")
else:
print("Zakljucak: Ne odbacujemo H0. Linearna povezanost nije statisticki znacajna.")
```

U gornjem primjeru reziduali su slučajno raspršeni oko osi nula, što znači da model nema sustavnu pogrešku i da je pretpostavka homoskedastičnosti zadovoljena. Ne vidimo jasan uzorak ni promjenu varijance uz rast predviđenih vrijednosti.

Zaključak: Budući da je p-vrijednost manja od 0.05, odbacujemo H_0 i zaključujemo da postoji statistički značajna linearna povezanost između razine ugljikovodika i čistoće kisika. To znači da nagib regresijskog pravca nije nula — promjena u x utječe na y. **Napomena**: "značajna" se odnosi na statističku potvrdu odnosa, a ne nužno na njegovu veličinu.

3 Zadaci za LV3

Tijekom laboratorijske vježbe na github-u će biti dostupni odgovarajući skupovi podataka za svaki zadatak.

Cilj vježbe nije puko izvođenje izračuna, već razumijevanje postupka statističkog zaključivanja, od provjere pretpostavki (normalnost, homoskedastičnost) do interpretacije rezultata testa ili modela.

Uz praktične zadatke, bit će potrebno **odgovoriti i na odgovarajuća pitanja** za teorijsko povezivanje praktičnog dijela te poticanje kritičkog zaključivanja. Pitanja služe za provjeru razumijevanja koncepata, objašnjenje dobivenih rezultata i procjenu njihove statističke i stvarne (praktične) važnosti. Pitanaj će također biti dostupna na github-u.

Za pripremu laboratorijske vježbe je nužno:

- proći gradivo s predavanja koje obuhvaća statističko zaključivanje o dvjema varijablama i osnovne pojmove regresijske analize,
- proučiti teorijski uvod u ovom predlošku i razumjeti osnovne testove (t-test, Wilcoxon, Mann–Whitney, linearna regresija),
- pregledati i proći Python notebook s predavanja radi upoznavanja sintakse i načina interpretacije rezultata.

Zadatak 1: Zaključivanje o nezavisnim uzorcima

- Podaci dvije grupe (dvije nezavisne varijable)
- Koraci:
  - Deskriptivna statistika i vizualizacija.
  - Provjera homoskedastičnosti (Leveneov test).
  - Postavljanje hipoteza.
  - Izbor i provođenje t-testa (Pooled ili Welchov).
  - Testiranje na dvije razine značajnosti ($\alpha = 0.05$, $\alpha = 0.01$).

Zadatak 2: Zaključivanje o zavisnim uzorcima

Opis: Testiranje razlika prije i poslije intervencije.

- Koriste se podaci prije i poslije intervencije.

- Provjerava se normalnost razlika (Shapiro-Wilk).
- Odabire se Paired t-test ili Wilcoxon signed-rank test.

Zadatak 3: Jednostavna linearna regresija

- Modeliranje odnosa $y = \beta_0 + \beta_1 x + \varepsilon$
- Procjena koeficijenata metodom najmanjih kvadrata
- interpretacija R^2 i testa značajnosti nagiba
- Analiza reziduala (vizualno i statistički)
- Predviđanje i intervali (CI i PI)

4 Zadaci - LV3 2025./2026.

Tijekom laboratorijske vježbe raditi s odabranim varijablama iz CalCOFI baze:

- `Depthm (dubina, m)`,
- `T_degC (temperatura mora, `^{o}`C)`,
- `Salnty (salinitet, PSU)`.

```
import kagglehub
import pandas as pd
# Download latest version
path = kagglehub.dataset_download("sohier/calcofi")
print("Path to dataset files:", path)
df = pd.read_csv(f"{path}/bottle.csv")
```

Cilj je provesti metode statističkog zaključivanja i jednostavne linearne regresije te analizirati smisao odnosa između ovih varijabli.

Zadatak 1: Statističko zaključivanje o temperaturi u dvjema dubinskim zonama

Cilj: Ispitati postoji li značajna razlika u temperaturi mora između površinskog i dubljeg sloja.

1. Podijelite podatke u dvije grupe:

   Grupa 1: $Depthm < 50$ m, Grupa 2: $Depthm > 200$ m.

2. Koristiti uzorak od 5000 slučajnih vrijednosti - za potrebe laboratorijske vježbe, jer je u suprotnom potrebno koristiti Kolmogorov-Smirnov teste za ispitivanje normalnosti.

```
shallow_sample = shallow.sample(5000, random_state=1)
deep_sample = deep.sample(5000, random_state=1)
```

3. Ispitajte normalnost distribucije (*Shapiro–Wilk test*) i jednakost varijanci (*Leveneov test*).

4. Ako su pretpostavke zadovoljene, koristite **t-test za nezavisne uzorke**; u suprotnom, koristite **Mann–Whitney U test**.

5. Testirajte hipoteze:

$$H_0 : \mu_{\text{pliće}} = \mu_{\text{dublje}}, \qquad H_1 : \mu_{\text{pliće}} \neq \mu_{\text{dublje}}.$$

6. Interpretirajte rezultat i komentirajte značenje (temperatura se obično smanjuje s dubinom).

Python primjer:

```
    shallow = df[df['Depthm'] < 50]['T_degC']
    deep = df[df['Depthm'] > 200]['T_degC']

    # Provjera normalnosti
    print(stats.shapiro(shallow))
    print(stats.shapiro(deep))

    # Leveneov test (jednakost varijanci)
    print(stats.levene(shallow, deep))

    # Usporedba - t-test ili Mann-Whitney
    t, p = stats.ttest_ind(shallow, deep, equal_var=False)
    print("t =", t, "p =", p)
    # ili:
    u, p = stats.mannwhitneyu(shallow, deep)
    print("U =", u, "p =", p)
```

Zadatak 2: Jednostavna linearna regresija (Temperatura – Dubina)

Cilj: Modelirati ovisnost temperature o dubini.

1. Napravite scatter dijagram s `Depthm` (x) i `T_degC` (y). Opisati tumačenje *scatter* dijagrama.

2. Procijenite model:

$$T\_degC = \beta_0 + \beta_1 \times Depthm + \varepsilon$$

   koristeći *linregress()* funkciju.

3. Izračunajte i interpretirajte:

   - koeficijente β_0 i β_1,

- koeficijent korelacije r i determinacije R^2,
- p-vrijednost za hipotezu $H_0 : \beta_1 = 0$.

4. Nacrtajte regresijski pravac i analizirajte reziduale (homoskedastičnost).

Primjer koda:

```
from scipy.stats import linregress
import matplotlib.pyplot as plt

x = df['Depthm']
y = df['T_degC']
slope, intercept, r, p, se = linregress(x, y)

print(f"y = {intercept:.2f} + {slope:.4f}x, R^2 = {r**2:.3f}")
plt.scatter(x, y, alpha=0.5, label="Podaci")
plt.plot(x, intercept + slope*x, color="red", label="
    Regresijski pravac")
plt.xlabel("Dubina (m)")
plt.ylabel("Temperatura ($^o$ C)")
plt.legend()
plt.show()
```

Napomena – priprema podataka za regresijsku analizu

Prije provođenja regresijske analize potrebno je ukloniti ekstremne i nerealne vrijednosti, jer **outlieri značajno utječu na nagib regresijskog pravca i koeficijent determinacije** R^2. Preporučuje se korištenje smislenog raspona dubina i temperatura te primjena **IQR metode (interkvartilnog raspona)** za filtriranje.
U nastavku je primjer koda koji prikazuje kako se podaci mogu pripremiti i očistiti prije regresijske analize:

```

# Ciscenje podataka prije regresije

import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import linregress

# 1) Smisleni rasponi za CalCOFI podatke
df_reg = df[
(df['Depthm'].between(0, 1000)) &
(df['T_degC'].between(-2, 30))
].copy()

print("Broj redaka prije ciscenja:", len(df))
print("Broj redaka nakon filtriranja:", len(df_reg))

# 2) Uklanjanje outliera pomocu IQR metode
Q1 = df_reg['T_degC'].quantile(0.25)
Q3 = df_reg['T_degC'].quantile(0.75)
IQR = Q3 - Q1
```

```
    lower = Q1 - 1.5 * IQR
    upper = Q3 + 1.5 * IQR

    df_reg = df_reg[
    (df_reg['T_degC'] >= lower) & (df_reg['T_degC'] <= upper)
    ]

    print("Broj redaka nakon IQR filtriranja:", len(df_reg))
```

Tumačenje: Nakon čišćenja podataka regresijski model bolje opisuje stvarni odnos između dubine i temperature. Uklanjanjem ekstremnih vrijednosti smanjuje se utjecaj pogrešnih mjerenja i postiže realniji nagib regresijskog pravca. Koeficijent determinacije R^2 pouzdanije pokazuje koliko promjena temperature objašnjava promjene dubine.

Zadatak 3: Usporedba rezultata prije i poslije intervencije

U istraživanju su sudjelovali ispitanici koji su prošli specifični program/intervenciju za poboljšanje vještina. Prikupljeni su rezultati testova prije (`Pre`) i poslije (`Post`) sudjelovanja u programu. Cilj je provjeriti postoji li statistički značajna razlika između rezultata prije i nakon intervencije.

1. **Analiza podataka:**
   - Učitajte priloženi skup podataka iz Excel tablice.
   - Provjerite osnovne statistike rezultata prije i poslije (*medijan, kvartile, raspon, srednju vrijednost, standardnu devijaciju*) – **deskriptivna statistika**.
2. **Odabir metode:**
   - Odredite jesu li podaci normalno distribuirani (npr. pomoću *Shapiro–Wilk testa*).
   - Ovisno o normalnosti razlika, odaberite odgovarajući test:
     - Ako su razlike normalne $\rightarrow$ koristite **parni t-test (Paired t-test)**.
     - Ako razlike nisu normalne $\rightarrow$ koristite **Wilcoxonov test s predznakom (Wilcoxon signed-rank test)**.
3. **Vizualizacija podataka:**
   - Prikazati podatke pomoću **boxplot** dijagrama za rezultate prije i poslije.
   - Dodatno, možete prikazati **histogram razlika (Pre–Post)** radi vizualne procjene distribucije.
4. **Interpretacija rezultata:**
   - Napišite kratki izvještaj s odgovorom na pitanje: *Postoji li statistički značajna razlika između rezultata prije i poslije intervencije?*
   - Uključite p-vrijednost, odabranu razinu značajnosti (α) i zaključak o odbacivanju ili neodbacivanju nulte hipoteze.