Palautettava tehtävä 8

June 1, 2025

1 Tehtävä 8: Merkitsevyystestaus

Tässä notebookissa: 1. Ladataan `Opinnäytetyökysely.xlsx` pandas-DataFrameksi. 2. **Tehtävä 1**: t-test (kulttuuri vs. tekniikka) muuttujalla *"Opinnäytetyön tekemisaika työviikkoina (40 h) aihekuvauksen tekemisestä työn valmistumiseen: työviikkoa"*. 3. **Tehtävä 2**: Mann-Whitney U-testi (kulttuuri vs. liiketalous) muuttujalla *"Hyödyllisyys: Seminaarit"*. 4. Tulostetaan ryhmien keskiarvot, 95 % luottamusvälit (tehtävä 1), Levene-testin ja t-testin tulokset sekä U-testin tulokset. 5. Esitetään nollahypoteesit sanatyyppisesti ja kirjoitetaan johtopäätökset (= 0.05).

1.1 1) Kirjastojen tuonti ja Excel-datan lataus

Tuodaan pandas ja scipy.stats. Luetaan sheet 0 (oletuksena "Kysely") DataFrameen.
Varmista, että tiedosto `Opinnäytetyökysely.xlsx` on samassa kansiossa kuin tämä notebook.

[9]:
```
import pandas as pd
import scipy.stats as stats

# Ladataan data
excel_path = r"D:\GitHub\PythonDataAnalytics\doc\Opinnäytetyökysely.xlsx"
df = pd.read_excel(excel_path, sheet_name=0, engine="openpyxl")

# Näytetään sarakenimet
df.columns.tolist()
```

[9]:
```
['Aikaleima',
 'Kuinka löysit aiheesi?',
 'Opinnäytetyöni oli hankkeistettu',
 'Oliko työsi teoreettinen vai käytännöllinen? Teoreettinen (1) - Käytännöllinen
(5)',
 'Pystyin itse vaikuttamaan aiheen valintaan',
 'Olin innostunut opinnäytetyötä tehdessäni',
 'Pystyin itse vaikuttamaan opinnäytetyöni ohjaajan valintaan',
 'Sain riittävästi ohjausta',
 'Hankin itse aktiivisesti tietoa työni aiheesta',
 'Tutkimusaiheeni kiinnosti minua',
 'Sain muilta opiskelijoilta tukea työni tekemisessä',
 'Työn toimeksiantaja oli kiinnostunut työstäni',
```

```
 'Työn toimeksiantaja oli kiinnostunut ohjaamaan työtäni',
 'Ohjaajani panos tuki työtäni',
 'Saamani ohjaus oli asiantuntevaa',
 'Valmistauduin ohjauspalavereihin',
 'Saamani ohjaus oli motivoivaa',
 'Työni ohjaaja vastasi nopeasti tiedusteluihini',
 'Ohjaustilanteet eivät tuntuneet minusta pelottavilta',
 'Ohjaajaani oli helppo lähestyä',
 'Luotin ohjaajani neuvoihin',
 'Eri ohjaustapojen käyttäminen ja niiden hyödyllisyys: Seminaarit',
 'Eri ohjaustapojen käyttäminen ja niiden hyödyllisyys: Henkilökohtaiset
tapaamiset ohjaajan kanssa',
 'Eri ohjaustapojen käyttäminen ja niiden hyödyllisyys: Sähköiset
yhteydet\xa0ohjaajan kanssa',
 'Eri ohjaustapojen käyttäminen ja niiden hyödyllisyys: Keskustelut
toimeksiantajan kanssa',
 'Hyödyllisyys: Seminaarit',
 'Hyödyllisyys:Henkilökohtaiset tapaamiset ohjaajan kanssa',
 'Hyödyllisyys: Sähköiset yhteydet\xa0ohjaajan kanssa',
 'Hyödyllisyys: Keskustelut toimeksiantajan kanssa',
 'Ohjaajan tuki opinnäytetyön eri vaiheissa:aiheen valinnassa',
 'Ohjaajan tuki opinnäytetyön eri vaiheissa:tutkimuskysymysten tai
kehittämistehtävän rajauksessa',
 'Ohjaajan tuki opinnäytetyön eri vaiheissa:menetelmien valinnassa',
 'Ohjaajan tuki opinnäytetyön eri vaiheissa:aineiston analyysissa',
 'Ohjaajan tuki opinnäytetyön eri vaiheissa:johtopäätösten ja yhteenvedon
tekemisessa',
 'Ohjaajan tuki opinnäytetyön eri vaiheissa:raportoinnissa',
 'Thesis grade',
 'Opinnäytetyön tekemiseen kulunut aika ensimmäisistä aihekaavailuista työn
valmistumiseen:kuukautta',
 'Opinnäytetyön tekemisaika työviikkoina (40 h) aihekuvauksen tekemisestä työn
valmistumiseen:työviikkoa',
 'Oliko työ parityö?',
 'Opiskeluala',
 'Sukupuoli',
 'Ikä']
```

1.2 2) Tehtävä 1: t-test (Kulttuuri vs. Tekniikka)

- Muuttuja:
  "Opinnäytetyön tekemisaika työviikkoina (40 h) aihekuvauksen tekemisestä työn valmistumiseen: työviikkoa".
- Ryhmät: Opiskeluala = "kulttuuri" vs. "tekniikka" (kirjoitetaan pienellä kirjaimella).
- Poistetaan rivit, joissa jomman kumman muuttujan arvo puuttuu (`dropna`).
- Lasketaan kummankin ryhmän n, keskiarvo ja 95 % luottamusvälit (t-jakauman mukaisesti).
- Testataan varianssien homogeenisuus Levene-testillä (= 0.05). Jos p $<$ 0.05, equal_var=False.

- Ajetaan t-test (`ttest_ind`) ja tulostetaan testisuure t ja p-arvo.
- Nollahypoteesi: “Kulttuuri- ja tekniikka-ryhmien opinnäytetyön tekemiseen kulunut aika ei eroa.”
- Johtopäätökset $= 0.05$: jos $p < 0.05$, hylätään nollahypoteesi ja todetaan ryhmien eroavaisuus; muuten ei hylätä.

[10]:
```
# Tehdään t-testin funktio tässä solussa:

def tee_t_test(df: pd.DataFrame):
    sarake = (
        "Opinnäytetyön tekemisaika työviikkoina (40 h) aihekuvauksen tekemisestä työn "
        "valmistumiseen:työviikkoa"
    )
    alaryhma = "Opiskeluala"

    # Poistetaan puuttuvat arvot:
    df1 = df.dropna(subset=[sarake, alaryhma]).copy()

    # Erotellaan ryhmät pienellä kirjaimella:
    kult = df1[df1[alaryhma].str.lower() == "kulttuuri"][sarake].astype(float)
    teki = df1[df1[alaryhma].str.lower() == "tekniikka"][sarake].astype(float)

    # Funktio 95 % CI:n laskemiseen
    def laske_95_ci(series):
        n = len(series)
        mean = series.mean()
        sem = stats.sem(series, nan_policy='omit')
        t_95 = stats.t.ppf(0.975, df=n - 1)
        delta = t_95 * sem
        return mean, mean - delta, mean + delta

    kult_mean, kult_ci_low, kult_ci_high = laske_95_ci(kult)
    teki_mean, teki_ci_low, teki_ci_high = laske_95_ci(teki)

    print("=== TEHTÄVÄ 1: t-test (Kulttuuri vs. Tekniikka) ===\n")
    print(
        f"Kulttuuri: n = {len(kult)}, "
        f"keskiarvo = {kult_mean:.2f}, 95 % CI = ({kult_ci_low:.2f}, {kult_ci_high:.2f})"
    )
    print(
        f"Tekniikka: n = {len(teki)}, "
        f"keskiarvo = {teki_mean:.2f}, 95 % CI = ({teki_ci_low:.2f}, {teki_ci_high:.2f})\n"
    )
```

```
    # Levene-test varianssien homogeenisuuden tarkastamiseen
    levene_stat, levene_p = stats.levene(kult, teki, center='mean')
    print(f"Levene-test: stat = {levene_stat:.4f}, p = {levene_p:.4f}")

    if levene_p < 0.05:
        print("→ Varianssit eivät ole homogeeniset (p < 0.05). Käytetään␣
↪equal_var=False.\n")
        equal_var = False
    else:
        print("→ Varianssit ovat homogeeniset (p  0.05). Käytetään␣
↪equal_var=True.\n")
        equal_var = True

    # t-test
    t_stat, p_val = stats.ttest_ind(kult, teki, equal_var=equal_var,␣
↪nan_policy='omit')
    print("t-test:")
    print(f"  testisuure t = {t_stat:.4f}, p = {p_val:.4f}")
    print(
        "  Nollahypoteesi: 'Kulttuuri- ja tekniikka-ryhmien opinnäytetyön␣
↪tekemisaika ei eroa.'\n"
    )

    if p_val < 0.05:
        print("  → P-arvo < 0.05, hylätään nollahypoteesi: ryhmien välillä on␣
↪tilastollisesti merkitsevä ero.")
    else:
        print("  → P-arvo  0.05, ei voida hylätä nollahypoteesia: ryhmien␣
↪välillä ei ole tilastollisesti merkitsevää eroa.")

    print("\nSanalliset johtopäätökset:")
    if p_val < 0.05:
        print(
            "- Kulttuuri- ja tekniikkaopiskelijoiden opinnäytetyön tekemiseen␣
↪kulunut aika "
            "eroaa tilastollisesti merkitsevästi (  = 0.05)."
        )
    else:
        print(
            "- Kulttuuri- ja tekniikkaopiskelijoiden opinnäytetyön tekemiseen␣
↪kulunut aika "
            "ei eroa tilastollisesti merkitsevästi (  = 0.05)."
        )
    print("\n" + "=" * 60 + "\n")

# Ajetaan t-test
```

```
tee_t_test(df)
```

```
=== TEHTÄVÄ 1: t-test (Kulttuuri vs. Tekniikka) ===

Kulttuuri: n = 100, keskiarvo = 22.36, 95 % CI = (18.93, 25.79)
Tekniikka: n = 50, keskiarvo = 14.28, 95 % CI = (11.97, 16.59)

Levene-test: stat = 13.8919, p = 0.0003
→ Varianssit eivät ole homogeeniset (p < 0.05). Käytetään equal_var=False.

t-test:
  testisuure t = 3.8913, p = 0.0002
  Nollahypoteesi: 'Kulttuuri- ja tekniikka-ryhmien opinnäytetyön tekemisaika ei
eroa.'

  → P-arvo < 0.05, hylätään nollahypoteesi: ryhmien välillä on tilastollisesti
merkitsevä ero.

Sanalliset johtopäätökset:
- Kulttuuri- ja tekniikkaopiskelijoiden opinnäytetyön tekemiseen kulunut aika
eroaa tilastollisesti merkitsevästi (  = 0.05).

============================================================
```

1.3 3) Tehtävä 2: Mann-Whitney U-testi (Kulttuuri vs. Liiketalous)

- Muuttuja: *"Hyödyllisyys: Seminaarit"*.

- Ryhmät: Opiskeluala = "kulttuuri" vs. "liiketalous" (pienellä kirjaimella).
- Poistetaan rivit, joissa seminaariarvio puuttuu (`dropna`).
- Lasketaan kummankin ryhmän n ja keskiarvo.
- Ajetaan Mann-Whitney U-testi (two-sided).

- Nollahypoteesi: "Kulttuuri- ja liiketalouden opiskelijoiden seminaarien hyödyllisyysarvioiden jakaumat eivät eroa."

- Johtopäätös = 0.05: jos p $<$ 0.05, hylätään nollahypoteesi; muuten ei hylätä.

[11]:
```
# Tehdään Mann-Whitney -funktiona tässä solussa:

def tee_mannwhitney(df: pd.DataFrame):
    sarake = "Hyödyllisyys: Seminaarit"
    alaryhma = "Opiskeluala"

    df2 = df.dropna(subset=[sarake, alaryhma]).copy()
    kult = df2[df2[alaryhma].str.lower() == "kulttuuri"][sarake].astype(float)
    liik = df2[df2[alaryhma].str.lower() == "liiketalous"][sarake].astype(float)
```

```
    print("=== TEHTÄVÄ 2: Mann-Whitney U (Kulttuuri vs. Liiketalous) ===\n")
    print(f"Kulttuuri: n = {len(kult)}, keskiarvo = {kult.mean():.2f}")
    print(f"Liiketalous: n = {len(liik)}, keskiarvo = {liik.mean():.2f}\n")

    u_stat, p_val = stats.mannwhitneyu(kult, liik, alternative='two-sided')
    print("Mann-Whitney U-testi:")
    print(f"  U = {u_stat:.4f}, p = {p_val:.4f}")
    print(
        "  Nollahypoteesi: "
        "'Kulttuuri- ja liiketalouden opiskelijoiden seminaarien hyödyllisyyden␣
↪arviot eivät eroa jakaumiltaan.'\n"
    )

    if p_val < 0.05:
        print("  → P-arvo < 0.05, hylätään nollahypoteesi: ryhmien välillä on␣
↪tilastollisesti merkitsevä ero.")
    else:
        print("  → P-arvo  0.05, ei voida hylätä nollahypoteesia: ryhmien␣
↪välillä ei ole tilastollisesti merkitsevää eroa.")

    print("\nSanalliset johtopäätökset:")
    if p_val < 0.05:
        print(
            "- Kulttuuri- ja liiketalouden opiskelijoiden arviot seminaarien␣
↪hyödyllisyydestä "
            "eroavat tilastollisesti merkitsevästi (  = 0.05)."
        )
    else:
        print(
            "- Kulttuuri- ja liiketalouden opiskelijoiden arviot seminaarien␣
↪hyödyllisyydestä "
            "eivät erotu tilastollisesti merkitsevästi (  = 0.05)."
        )
    print("\n" + "=" * 60 + "\n")

# Ajetaan Mann-Whitney -testi
tee_mannwhitney(df)
```

```
=== TEHTÄVÄ 2: Mann-Whitney U (Kulttuuri vs. Liiketalous) ===

Kulttuuri: n = 122, keskiarvo = 3.95
Liiketalous: n = 52, keskiarvo = 3.19

Mann-Whitney U-testi:
  U = 4436.0000, p = 0.0000
  Nollahypoteesi: 'Kulttuuri- ja liiketalouden opiskelijoiden seminaarien
```

```
hyödyllisyyden arviot eivät eroa jakaumiltaan.'

  → P-arvo < 0.05, hylätään nollahypoteesi: ryhmien välillä on tilastollisesti
merkitsevä ero.

Sanalliset johtopäätökset:
- Kulttuuri- ja liiketalouden opiskelijoiden arviot seminaarien hyödyllisyydestä
eroavat tilastollisesti merkitsevästi (  = 0.05).

============================================================
```

2 Yhteenveto

- **Tehtävä 1 (t-test):**
  - Keskiarvot ja 95 % luottamusvälit:
    - Kulttuuri: … (n, mean, CI)
    - Tekniikka: … (n, mean, CI)
  - Levene-test: stat, p
  - T-test: t, p
  - Nollahypoteesi: “Kulttuuri- ja tekniikka-ryhmien opinnäytetyön tekemisaika ei eroa.”
  - Johtopäätös: hylätään/ei hylätä nollahypoteesia (= 0.05)
  - Sanallinen johtopäätös lyhyesti (ryhmien välinen ero/ei eroa).
- **Tehtävä 2 (Mann-Whitney U):**
  - Keskiarvot:
    - Kulttuuri: … (n, mean)
    - Liiketalous: … (n, mean)
  - U-test: U, p
  - Nollahypoteesi: “Kulttuuri- ja liiketalouden opiskelijoiden seminaarien hyödyllisyysarviot eivät eroa jakaumiltaan.”
  - Johtopäätös: hylätään/ei hylätä nollahypoteesia (= 0.05)
  - Sanallinen johtopäätös lyhyesti (ryhmien välinen ero/ei eroa).