P5_Two_Sample_Hypothesis_Test

April 6, 2023

[ ]:
```
# Import Libraries
import pandas as pd
import scipy.stats as st
import statsmodels.stats.weightstats as ws
from statsmodels.stats.proportion import proportions_ztest

# Read csv file
df = pd.read_csv("../data/anggur.csv")

display(df)
```

```
     fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0             5.90            0.4451       0.1813        2.049401   0.070574
1             8.40            0.5768       0.2099        3.109590   0.101681
2             7.54            0.5918       0.3248        3.673744   0.072416
3             5.39            0.4201       0.3131        3.371815   0.072755
4             6.51            0.5675       0.1940        4.404723   0.066379
..             …               …            …             …          …
995           7.96            0.6046       0.2662        1.592048   0.057555
996           8.48            0.4080       0.2227        0.681955   0.051627
997           6.11            0.4841       0.3720        2.377267   0.042806
998           7.76            0.3590       0.3208        4.294486   0.098276
999           5.87            0.5214       0.1883        2.179490   0.052923

     free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0              16.593818                 42.27   0.9982  3.27       0.71
1              22.555519                 16.01   0.9960  3.35       0.57
2               9.316866                 35.52   0.9990  3.31       0.64
3              18.212300                 41.97   0.9945  3.34       0.55
4               9.360591                 46.27   0.9925  3.27       0.45
..                   …                     …      …   …          …
995            14.892445                 44.61   0.9975  3.35       0.54
996            23.548965                 25.83   0.9972  3.41       0.46
997            21.624585                 48.75   0.9928  3.23       0.55
998            12.746186                 44.53   0.9952  3.30       0.66
999            16.203864                 24.37   0.9983  3.29       0.70

     alcohol  quality
```

```
0       8.64        7
1      10.03        8
2       9.23        8
3      14.07        9
4      11.49        8
..       …       …
995    10.41        8
996     9.91        8
997     9.94        7
998     9.76        8
999    10.17        7

[1000 rows x 12 columns]
```

1 Pengujian Hipotesis Terhadap Dua Sampel

1.0.1 Langkah-Langkah Pembuktian Hipotesis:

1. Tentukan hipotesis nol H_0.
2. Tentukan hipotesis alternatif H_1.
3. Tentukan tingkat signifikan α.
4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.
5. Hitung nilai uji statistik dari data sample. Hitung *p-value* sesuai dengan uji statistik yang digunakan.
6. Ambil keputusan "Tolak H_0" jika nilai uji statistik terletak di daerah kritis, atau dengan tes signifikan, "Tolak H_0" jika *p-value* lebih kecil dibanding tingkat signifikansi α yang diinginkan.

1.1 Q1: Data kolom fixed acidity dibagi 2 sama rata: bagian awal dan bagian akhir kolom. Benarkah rata-rata kedua bagian tersebut sama?

Sampel pengujian: - sampel_1: bagian awal kolom 'fixed acidity' - sampel_2: bagian akhir kolom 'fixed acidity'

Langkah-langkah: 1. H_0: $\mu_1 - \mu_2 = 0$ (rata-rata kedua sampel sama) 2. H_1: $\mu_1 - \mu_2 \neq 0$ (rata-rata kedua sampel berbeda) 3. Penentuan tingkat signifikan: $\alpha = 0.05$ 4. Penentuan uji statistik dan daerah kritis: - Standar deviasi populasi (σ) dari kedua sampel diketahui sama karena diambil dari populasi yang sama - Uji hipotesis adalah *two-tailed test* - Oleh karena itu, rumus pengujian yang digunakan adalah sebagai berikut

$$z = \frac{(\bar{x_1} - \bar{x_2}) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

- Daerah kritis adalah $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$ 5. Perhitungan nilai uji statistik z ada pada kode di bawah ini. 6. Pengambilan keputusan: - Tolak H_0 jika $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$ - H_0 tidak ditolak jika $-z_{\alpha/2} \leq z \leq z_{\alpha/2}$

[ ]:
```
# Sample setup
fixed_acidity = df['fixed acidity']
fixed_acidity_sample_1 = fixed_acidity[:len(fixed_acidity)//2]
fixed_acidity_sample_2 = fixed_acidity[len(fixed_acidity)//2:]
```

```python
# Test statistic calculation
diff = 0
significance = 0.05

z_value, ztest_pvalue_1 = ws.ztest(fixed_acidity_sample_1,␣
 ↪fixed_acidity_sample_2, value=diff)

z_alpha_over_2 = st.norm.ppf(1 - significance/2)

# Drawing a conclusion
print(f"Critical region: z < {-z_alpha_over_2} or z > {z_alpha_over_2}")
print(f"Test statistic: z = {z_value}")
print(f"p-value = {ztest_pvalue_1}")
print()
if (z_value < -z_alpha_over_2 or z_value > z_alpha_over_2):
    print("Nilai z berada dalam critical region")
    verdict = "H0 ditolak, rata-rata sampel 1 tidak sama dengan rata-rata␣
 ↪sampel 2"
else:
    print("Nilai z berada di luar critical region")
    verdict = "H0 tidak ditolak, rata-rata sampel 1 sama dengan rata-rata␣
 ↪sampel 2"

if (ztest_pvalue_1 < significance):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0")

print("\nKesimpulan: " + verdict)
```

```
Critical region: z < -1.959963984540054 or z > 1.959963984540054
Test statistic: z = 0.02604106999906379
p-value = 0.9792245804254097

Nilai z berada di luar critical region
Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tidak tolak H0

Kesimpulan: H0 tidak ditolak, rata-rata sampel 1 sama dengan rata-rata sampel 2
```

1.2 Q2: Data kolom chlorides dibagi 2 sama rata: bagian awal dan bagian akhir kolom. Benarkah rata-rata bagian awal lebih besar daripada bagian akhir sebesar 0.001?

Sampel pengujian: - sampel_1: bagian awal kolom 'chlorides' - sampel_2: bagian akhir kolom 'chlorides'

Langkah-langkah: 1. H_0: $\mu_1 - \mu_2 = 0.001$ (rata-rata bagian awal lebih besar daripada bagian akhir sebesar 0.001) 2. H_1: $\mu_1 - \mu_2 > 0.001$ (rata-rata bagian awal lebih besar daripada bagian akhir dengan selisih lebih dari 0.001) 3. Penentuan tingkat signifikan: $\alpha = 0.05$ 4. Penentuan uji statistik dan daerah kritis: - Standar deviasi populasi (σ) dari kedua sampel diketahui sama karena diambil dari populasi yang sama - Uji hipotesis adalah *one-tailed test*, dengan *critical region* berada pada sisi kanan grafik distribusi nilai - Oleh karena itu, rumus pengujian yang digunakan adalah sebagai berikut

$$z = \frac{(\bar{x_1} - \bar{x_2}) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

- Daerah kritis adalah $z > z_\alpha$ 5. Perhitungan nilai uji statistik z ada pada kode di bawah ini. 6. Pengambilan keputusan: - Tolak H_0 jika $z > z_\alpha$ - H_0 tidak ditolak jika $z \leq z_\alpha$

[ ]:
```
# Sample setup
chlorides = df['chlorides']
chlorides_sample_1 = chlorides[:len(chlorides)//2]
chlorides_sample_2 = chlorides[len(chlorides)//2:]

# Test statistic calculation
diff = 0.001
significance = 0.05

z_value, ztest_pvalue_2 = ws.ztest(chlorides_sample_1, chlorides_sample_2,␣
 ↪value=diff)

z_alpha = st.norm.ppf(1 - significance)


# Drawing a conclusion
print(f"Critical region: z > {z_alpha}")
print(f"Test statistic: z = {z_value}")
print(f"p-value = {ztest_pvalue_2}")
print()
if (z_value > z_alpha):
    print("Nilai z berada dalam critical region")
    verdict = "H0 ditolak, rata-rata sampel 1 lebih besar dari rata-rata sampel␣
 ↪2, tetapi selisih lebih dari 0.001"
else:
    print("Nilai z berada di luar critical region")
    verdict = "H0 tidak ditolak, rata-rata sampel 1 lebih besar dari rata-rata␣
 ↪sampel 2 sebanyak 0.001"
```

```
if (ztest_pvalue_2 < significance):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0")


print("\nKesimpulan: " + verdict)
```

```
Critical region: z > 1.6448536269514722
Test statistic: z = -0.467317122852132
p-value = 0.640273007581107

Nilai z berada di luar critical region
Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tidak tolak H0

Kesimpulan: H0 tidak ditolak, rata-rata sampel 1 lebih besar dari rata-rata
sampel 2 sebanyak 0.001
```

1.3 Q3: Benarkah rata-rata sampel 25 baris pertama kolom Volatile Acidity sama dengan rata-rata 25 baris pertama kolom Sulphates ?

Sampel pengujian: - sampel_1: 25 baris pertama kolom 'volatile acidity' - sampel_2: 25 baris pertama kolom 'sulphates'

Langkah-langkah: 1. H_0: $\mu_1 - \mu_2 = 0$ (rata-rata kedua sampel sama) 2. H_1: $\mu_1 - \mu_2 \neq 0$ (rata-rata kedua sampel berbeda) 3. Penentuan tingkat signifikan: $\alpha = 0.05$ 4. Penentuan uji statistik dan daerah kritis: - Standar deviasi populasi (σ) dari kedua sampel diketahui berbeda - Uji hipotesis adalah *two-tailed test* - Oleh karena itu, rumus pengujian yang digunakan adalah sebagai berikut

$$t = \frac{(\bar{x_1} - \bar{x_2}) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

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

- Daerah kritis adalah $t < -t_{\alpha/2}$ atau $t > t_{\alpha/2}$ 5. Perhitungan nilai uji statistik t ada pada kode di bawah ini. 6. Pengambilan keputusan: - Tolak H_0 jika $t < -t_{\alpha/2}$ atau $t > t_{\alpha/2}$ - H_0 tidak ditolak jika $-t_{\alpha/2} \leq t \leq t_{\alpha/2}$

[ ]:
```
# Sample setup
volatile_acidity = df['volatile acidity']
sample_1_volatile_acidity = volatile_acidity[:25]

sulphates = df['sulphates']
sample_2_sulphates = sulphates[:25]

# Test statistic calculation
```

```
diff = 0
significance = 0.05

t_value, ttest_pvalue, dof = ws.ttest_ind(sample_1_volatile_acidity,␣
 ↪sample_2_sulphates, value=diff)

t_alpha_over_2 = st.t.ppf(1 - significance/2, dof)

# Drawing a conclusion
print(f"Critical region: t < {-t_alpha_over_2} or t > {t_alpha_over_2}")
print(f"Degree of Freedom: v = {dof}")
print(f"Test statistic: t = {t_value}")
print(f"p-value = {ttest_pvalue}")
print()
if (t_value < -t_alpha_over_2 or t_value > t_alpha_over_2):
    print("Nilai t berada dalam critical region")
    verdict = "H0 ditolak, rata-rata sampel 1 tidak sama dengan rata-rata␣
 ↪sampel 2"
else:
    print("Nilai t berada di luar critical region")
    verdict = "H0 tidak ditolak, rata-rata sampel 1 sama dengan rata-rata␣
 ↪sampel 2"

if (ttest_pvalue < significance):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0")


print("\nKesimpulan: " + verdict)
```

```
Critical region: t < -2.0106347546964454 or t > 2.0106347546964454
Degree of Freedom: v = 48.0
Test statistic: t = -2.6374821676748703
p-value = 0.011223058174680032

Nilai t berada dalam critical region
Nilai p lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tolak H0

Kesimpulan: H0 ditolak, rata-rata sampel 1 tidak sama dengan rata-rata sampel 2
```

1.4 Q4: Bagian awal kolom residual sugar memiliki variansi yang sama dengan bagian akhirnya?

Sampel pengujian: - sampel_1: bagian awal dari kolom 'residual sugar' - sampel_2: bagian akhir dari kolom 'residual sugar'

Langkah-langkah: 1. H_0: $\sigma_1^2 = \sigma_2^2$ (variansi kedua sampel sama) 2. H_1: $\sigma_1^2 \neq \sigma_2^2$ (variansi kedua sampel berbeda) 3. Penentuan tingkat signifikan: $\alpha = 0.05$ 4. Penentuan uji statistik dan daerah kritis: - Uji hipotesis adalah *two-tailed test* - Oleh karena itu, rumus pengujian yang digunakan adalah sebagai berikut

$$f = \frac{s_1^2}{s_2^2}$$

- Daerah kritis adalah $f < f_{1-\alpha/2}(v_1, v_2)$ atau $f > f_{\alpha/2}(v_1, v_2)$ 5. Perhitungan nilai uji statistik f ada pada kode di bawah ini. 6. Pengambilan keputusan: - Tolak H_0 jika $f < f_{1-\alpha/2}(v_1, v_2)$ atau $f > f_{\alpha/2}(v_1, v_2)$ - H_0 tidak ditolak jika $f_{1-\alpha/2}(v_1, v_2) \leq f \leq f_{\alpha/2}(v_1, v_2)$

[ ]:
```
# Sample setup
residual_sugar = df['chlorides']
residual_sugar_sample_1 = residual_sugar[:len(residual_sugar)//2]
residual_sugar_sample_2 = residual_sugar[len(residual_sugar)//2:]

# Hypothesis testing setup
sample_1_variance = st.variation(residual_sugar_sample_1, ddof=1)
sample_2_variance = st.variation(residual_sugar_sample_2, ddof=1)
print(f"Sample_1 variance: s1^2 = {sample_1_variance}")
print(f"Sample_2 variance: s2^2 = {sample_2_variance}")
print()

# Test statistic calculation
diff = 0
significance = 0.05

f_value = sample_1_variance / sample_2_variance


# f-distribution test critical points, note: ppf accepts left-side percentage
f_left_tail = st.f.ppf(1-(1 - significance/2), len(residual_sugar_sample_1)-1,␣
 ↪len(residual_sugar_sample_2)-1)
f_right_tail = st.f.ppf(1-(significance/2), len(residual_sugar_sample_1)-1,␣
 ↪len(residual_sugar_sample_2)-1)
f_test_pvalue = st.f.cdf(f_value, len(residual_sugar_sample_1)-1,␣
 ↪len(residual_sugar_sample_2)-1)

# Drawing a conclusion
print(f"Critical region: f < {f_left_tail} or f > {f_right_tail}")
print(f"Test statistic: f = {f_value}")
print(f"p-value = {f_test_pvalue}")
print()
```

```
if (f_value < f_left_tail or f_value > f_right_tail):
    print("Nilai f berada dalam critical region")
    verdict = "H0 ditolak, variansi kedua sampel berbeda"
else:
    print("Nilai f berada di luar critical region")
    verdict = "H0 tidak ditolak, variansi kedua sampel sama"

if (f_test_pvalue < significance):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0")


print("\nKesimpulan: " + verdict)
```

```
Sample_1 variance: s1^2 = 0.24774799282054896
Sample_2 variance: s2^2 = 0.24783954608585196

Critical region: f < 0.8388857772763105 or f > 1.1920574017201653
Test statistic: f = 0.9996305946054659
p-value = 0.49835451097845074

Nilai f berada di luar critical region
Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tidak tolak H0

Kesimpulan: H0 tidak ditolak, variansi kedua sampel sama
```

1.5 Q5: Proporsi nilai setengah bagian awal alcohol yang lebih dari 7, adalah lebih besar daripada, proporsi nilai yang sama di setengah bagian akhir alcohol?

Sampel pengujian: - sampel_1: bagian awal dari kolom 'alcohol' yang bernilai lebih dari 7 - sampel_2: bagian akhir dari kolom 'alcohol' yang bernilai lebih dari 7

Langkah-langkah: 1. H_0: $p_1 - p_2 = 0$ (proporsi kedua sampel sama) 2. H_1: $p_1 - p_2 > 0$ (proporsi sampel pertama lebih besar dari proporsi sampel kedua) 3. Penentuan tingkat signifikan: $\alpha = 0.05$ 4. Penentuan uji statistik dan daerah kritis: - Uji hipotesis adalah *one-tailed test* - Oleh karena itu, rumus pengujian yang digunakan adalah sebagai berikut

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

- Daerah kritis adalah $z > z_\alpha$ 5. Perhitungan nilai uji statistik z ada pada kode di bawah ini. 6. Pengambilan keputusan: - Tolak H_0 jika $z > z_\alpha$ - H_0 tidak ditolak jika $z \leq z_\alpha$

[ ]:
```
# Sample setup
alcohol = df['alcohol']
alcohol_sample_1 = alcohol[:len(alcohol)//2]
alcohol_sample_2 = alcohol[len(alcohol)//2:]

# Filter sample to greater than 7
alcohol_sample_1_gt7 = alcohol_sample_1[alcohol_sample_1 > 7]
alcohol_sample_2_gt7 = alcohol_sample_2[alcohol_sample_2 > 7]

# Hypothesis testing setup
x1_x2 = [len(alcohol_sample_1_gt7), len(alcohol_sample_2_gt7)]
n1_n2 = [len(alcohol_sample_1), len(alcohol_sample_2)]
print(f"x1, x2 = {x1_x2}")
print(f"n1, n2 = {n1_n2}")

# Test statistic calculation
diff = 0
significance = 0.05
stat, proportion_ztest_pvalue = proportions_ztest(x1_x2, n1_n2, value=diff)

z_alpha = st.norm.ppf(1 - significance)


# Drawing a conclusion
print(f"Critical region: z > {z_alpha}")
print(f"Test statistic: z = {z_value}")
print(f"p-value = {proportion_ztest_pvalue}")
print()
if (z_value > z_alpha):
    print("Nilai z berada dalam critical region")
    verdict = "H0 ditolak, proporsi sampel 1 lebih besar dari proporsi sampel 2"
else:
    print("Nilai z berada di luar critical region")
    verdict = "H0 tidak ditolak, proporsi sampel 1 sama dengan proporsi sampel␣
 ↪2"

if (proportion_ztest_pvalue < significance):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0")


print("\nKesimpulan: " + verdict)
```

```
x1, x2 = [495, 495]
n1, n2 = [500, 500]
```

```
Critical region: z > 1.6448536269514722
Test statistic: z = -0.467317122852132
p-value = 1.0

Nilai z berada di luar critical region
Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tidak tolak H0

Kesimpulan: H0 tidak ditolak, proporsi sampel 1 sama dengan proporsi sampel 2
```