Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung

Tugas Besar IF2220 Probabilitas dan Statistika

Penarikan Kesimpulan dan Pengujian Hipotesis

Tujuan:

- Mahasiswa memahami dan dapat menyelesaikan persoalan distribusi peluang variabel random diskrit dan kontinu, dan
- mahasiswa mampu menyelesaikan persoalan untuk menarik kesimpulan mengenai parameter populasi yang diperoleh dari data hasil eksperimen.
- Mahasiswa mampu menyelesaikan persoalan pengujian hipotesis.

Petunjuk pengerjaan tugas:

1. Dikerjakan berkelompok (2 orang) dalam kelas yang sama.
2. Untuk menjawab soal, mahasiswa diharuskan membuat program bahasa R/Python yang ditulis pada Jupyter Notebook.
3. Arsip yang dikumpulkan: **File zip** yang berisi file **.ipynb** dan **.pdf** hasil *export* dari notebook dengan nama file **[Kelas]-T1-IF2220-[NIM].zip** dengan NIM adalah NIM terkecil anggota kelompok dan Kelas adalah K01,K02, dan sebagainya. File zip dapat diunggah melalui pranala berikut.
4. Tuliskan nomor soal dan keterangan pengerjaan selengkap mungkin dengan menggunakan Text di Jupyter Notebook, atau di comment di badan Code.
5. Untuk tes hipotesis, wajib menuliskan ke-6 langkah testing.
6. Batas pengumpulan adalah **16 April 2022 pukul 23.59**.

Enam Langkah Testing:

1. Tentukan Hipotesis nol (H_0: $\theta = \theta_0$), dimana θ bisa berupa μ, σ^2, p, atau data lain berdistribusi tertentu (normal, binomial, dsc.).
2. Pilih hipotesis alternatif H_1 salah dari dari $\theta > \theta_0$, $\theta < \theta_0$, atau $\theta \neq \theta_0$.
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 dengan TOLAK H_0 jika nilai uji terletak di daerah kritis atau dengan tes signifikan, TOLAK H_0 jika *p-value* lebih kecil dibanding tingkat signifikansi α yang diinginkan.

Soal Tugas

Diberikan sebuah data water_potability.csv yang dapat diakses pada utas berikut: Dataset Tugas Besar IF2220. water_potability.csv merupakan data metrik kualitas air yang mengandung 11 kolom sebagai berikut:

1. id
2. pH
3. Hardness
4. Solids
5. Chloramines
6. Sulfate
7. Conductivity
8. OrganicCarbon
9. Trihalomethanes
10. Turbidity
11. Potability

Kolom 2-10 adalah kolom atribut (non-target), sedangkan kolom 11 adalah kolom target.

Anda diminta untuk melakukan analisis statistika sebagai berikut:

1. Menulis deskripsi statistika (*Descriptive Statistics*) dari semua kolom pada data yang bersifat numerik, terdiri dari mean, median, modus, standar deviasi, variansi, range, nilai minimum, maksimum, kuartil, IQR, skewness dan kurtosis. Boleh juga ditambahkan deskripsi lain.
2. Membuat Visualisasi plot distribusi, dalam bentuk histogram dan boxplot untuk setiap kolom numerik. Berikan uraian penjelasan kondisi setiap kolom berdasarkan kedua plot tersebut.
3. Menentukan setiap kolom numerik berdistribusi normal atau tidak. Gunakan normality test yang dikaitkan dengan histogram plot.
4. Melakukan test hipotesis 1 sampel, dengan menuliskan 6 langkah testing dan menampilkan juga boxplotnya untuk kolom/bagian yang bersesuaian.

a. Nilai Rata-rata pH di atas 7?
b. Nilai Rata-rata Hardness tidak sama dengan 205?
c. Nilai Rata-rata 100 baris pertama kolom Solids bukan 21900?
d. Proporsi nilai Conductivity yang lebih dari 450, adalah tidak sama dengan 10%?
e. Proporsi nilai Trihalomethanes yang kurang dari 40, adalah kurang dari 5%?

5. Melakukan test hipotesis 2 sampel, dengan menuliskan 6 langkah testing dan menampilkan juga boxplotnya untuk kolom/bagian yang bersesuaian.
   a. Data kolom Sulfate dibagi 2 sama rata: bagian awal dan bagian akhir kolom. Benarkah rata-rata kedua bagian tersebut sama?
   b. Data kolom OrganicCarbon dibagi 2 sama rata: bagian awal dan bagian akhir kolom. Benarkah rata-rata bagian awal lebih besar dari pada bagian akhir sebesar 0.15?
   c. Rata-rata 100 baris pertama kolom Chloramines sama dengan 100 baris terakhirnya?
   d. Proporsi nilai bagian awal Turbidity yang lebih dari 4, adalah lebih besar daripada, proporsi nilai yang sama di bagian akhir Turbidity ?
   e. Bagian awal kolom Sulfate memiliki variansi yang sama dengan bagian akhirnya?
6. Test korelasi: tentukan apakah setiap kolom non-target berkorelasi dengan kolom target, dengan menggambarkan juga scatter plot nya. Gunakan correlation test.

Komponen Penilaian:

- Nomor 1 dan 2 : Kelengkapan jawaban dan ketepatan nilai
- Nomor 3, 4, 5, dan 6 : Kelengkapan jawaban, ketepatan nilai, dan kejelasan metode yang digunakan

Lain-lain:

1. Keterlambatan pengumpulan akan menyebabkan nilai menjadi nol.
2. Segala bentuk kecurangan akan ditindaklanjuti oleh asisten.
3. Segala pertanyaan hanya dapat ditanyakan melalui pranala berikut.

Referensi:

1. Dokumentasi R - https://www.rdocumentation.org/
2. Project Jupyter - http://jupyter.org/
3. Pandas - https://pandas.pydata.org/
4. Matplotlib - https://matplotlib.org/
5. Walpole, dkk. 2012. Probability and Statistics for Engineers and Scientists: Ninth Edition.