

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. Tugas terdiri dari 5 dataset berbeda yang dapat dipilih oleh masing-masing kelompok mahasiswa. Kumpulan dataset dapat diunduh melalui link folder Google Drive berikut [Dataset](#).
2. Tugas terdiri dari 6 buah soal. Terdapat 4 buah soal general yang bersifat sama untuk setiap jenis dataset. Terdapat juga 2 buah soal spesifik yang khusus pada dataset tersebut.
3. Dalam implementasi, lakukanlah perhitungan dengan menggunakan fungsi yang dibuat sendiri. Bandingkan hasilnya dengan fungsi yang tersedia dari library statistika seperti *scipy*.
4. Dikerjakan berkelompok (2 orang) dalam kelas yang sama. Pengisian kelompok dapat dilakukan pada sheets berikut [Pembagian Kelompok](#). Deadline pengisian pembagian kelompok adalah **Rabu, 8 Mei 2024, pukul 18:00**.
5. Untuk menjawab soal, mahasiswa diharuskan membuat program bahasa Python yang ditulis pada Jupyter Notebook.
6. Arsip yang dikumpulkan: **File zip** yang berisi file **.ipynb** dan **.pdf** hasil *export* dari notebook dengan nama file **[Kelas]-IF2220-[NIM1]-[NIM2].zip** dengan NIM1 adalah NIM terkecil anggota kelompok dan Kelas adalah K01, K02, dan sebagainya.
7. Pengumpulan tugas dilakukan melalui Edunex.
8. Tuliskan nomor soal dan keterangan pengerjaan selengkap mungkin dengan menggunakan badan teks di Jupyter Notebook.
9. Untuk tes hipotesis, wajib menuliskan ke-6 langkah testing.
10. Batas pengumpulan adalah **Jumat, 24 Mei 2024, 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 $\theta > \theta_0$, $\theta < \theta_0$, 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 General [Untuk setiap jenis dataset]:

1. Menulis deskripsi statistika (Descriptive Statistics) dari semua kolom pada data. Data yang bersifat numerik dapat diberikan nilai mean, median, modus, standar deviasi, variansi, range, nilai minimum, maksimum, kuartil, IQR, skewness dan kurtosis. Data dalam bentuk string dapat dicari unique values, dan proporsi nya.
2. Apakah pada data tersebut terdapat outlier? Jika ya, dapatkah anda menanganinya? Jelaskan apa yang umumnya dilakukan untuk menangani outlier.
3. Membuat Visualisasi plot distribusi. Berikan uraian penjelasan kondisi setiap kolom berdasarkan kedua plot tersebut. Jika numerik dapat dibuat dalam bentuk histogram dan box plot, dan jika string dengan histogram.
4. Menentukan distribusi setiap kolom numerik menggunakan hasil visualisasi histogram. Apakah kolom tersebut berdistribusi normal? Jika bukan, terdistribusi seperti apa kolom tersebut?

Dataset Banana

Seorang mahasiswa tingkat 3 sedang menjalani kerja praktek di sebuah perusahaan yang bergerak di industri buah-buahan. Perusahaan tersebut menghadapi kekhawatiran terkait kualitas buah pisang yang diberikan oleh pemasoknya. Sebagai bagian dari tugasnya, mahasiswa tersebut diberikan sebuah dataset yang berisikan informasi dan atribut-atribut yang terkait dengan buah pisang yang diberikan oleh pemasok. Mahasiswa diminta untuk melakukan analisis statistika terhadap dataset tersebut guna membantu perusahaan dalam memahami kualitas buah pisang yang mereka terima serta membantu dalam melakukan berbagai pengujian berbagai hipotesis.

Atribut: Acidity, Weight, Length, Appearance, Tannin, Ripeness, Sweetness, Country_of_Origin, Firmness, Grade, Price

Gunakan $\alpha = 0.05$

5. Hipotesis 1 sampel

- Perusahaan menerima beberapa keluhan bahwa buah pisang yang mereka terima akhir-akhir ini cukup asam. Dapatkah anda mengecek apakah rata-rata nilai Acidity di atas 6?
- Supplier menjanjikan bahwa rata-rata berat buah pisang adalah 150 gram. Pemilik mencurigai kebenaran hal ini. Apakah rata-rata buah pisang yang mereka kirim tidak bernilai 150 gram?
- Periksa apakah rata-rata panjang buah pisang 10 baris terakhir tidak sama dengan 49!
- Apakah proporsi nilai Tannin yang lebih besar dari 8 tidak sama dengan 55% dari total dataset?

6. Hipotesis 2 sampel

Perusahaan ingin membandingkan kualitas buah yang diterima pada paruh awal dan paruh akhir kerjasama. Anda dapat melakukan ini dengan membagi 1 dataset menjadi 2 bagian yang sama panjang.

- Anda diminta untuk memeriksa apakah rata-rata acidity dari buah pisang yang disuplai bernilai sama pada kedua kurun waktu tersebut.
- Bandingkanlah rata-rata appearance pada bagian awal dan akhir. Apakah rata-rata appearance pada dataset bagian awal lebih besar daripada bagian akhir sebesar 0.1 unit?
- Apakah variansi dari panjang pisang yang dipasok supplier sama pada bagian awal dan akhir?

- Apakah proporsi berat pisang yang lebih dari 150 pada dataset awal lebih besar daripada proporsi di bagian dataset akhir?

Dataset Candy:

Andi baru saja mendapatkan pekerjaan di sebuah perusahaan permen. Di hari pertama kerja, ia diminta untuk mengamati permen-permen yang dihasilkan pada pabrik. Selain itu, ia juga diberikan dataset yang berisi informasi mengenai permen-permen yang baru diproduksi. Lalu, ia diminta untuk melakukan analisis statistika terhadap permen-permen tersebut serta melakukan berbagai pengujian terhadap berbagai hipotesis. Bantulah Andi dalam melakukan hal-hal tersebut!

Atribut: Calories, Serving, Protein, Sugar, Sodium, Fat, Fiber, Flavour, Popularity

Gunakan $\alpha = 0.05$

5. Hipotesis 1 sampel

- Perusahaan menerima beberapa keluhan bahwa permennya kurang manis. Periksa! apakah anda mengecek apakah rata-rata nilai Sugar di bawah 25?
- Pada umumnya, rata-rata Serving untuk permen adalah 40 gram. Oleh karena itu, periksa! apakah rata-rata Serving permen yang diproduksi tidak bernilai 40 gram!
- Periksa! apakah rata-rata Sodium untuk permen 20 baris terakhir tidak sama dengan 74!
- Periksa! apakah proporsi nilai Protein yang lebih besar dari 3 tidak sama dengan 60% dari total dataset!

6. Hipotesis 2 sampel

Perusahaan ingin membandingkan kualitas permen yang diproduksi pada paruh awal dan paruh akhir produksi. Hal ini dapat dilakukan dengan membagi 1 dataset menjadi 2 bagian yang sama panjang.

- Periksa! apakah rata-rata Sugar dari permen yang diproduksi bernilai sama pada kedua kurun waktu!
- Bandingkan rata-rata Protein dari permen pada paruh awal dan akhir. Apakah rata-rata Protein pada dataset bagian awal lebih besar daripada bagian akhir sebesar 0.3 unit?
- Periksa! apakah variansi dari Sodium dari permen sama pada paruh awal dan akhir!
- Periksa! apakah proporsi Calories dari permen yang lebih dari 200 pada paruh awal lebih besar daripada proporsi di paruh akhir!

Dataset Weather:

Doni adalah seorang ahli meteorologi yang bekerja di sebuah perusahaan penyedia informasi cuaca teratur, yang bernama BMKG (Badan Meteorologi Klimatologi dan Geofisika). Perusahaan tersebut menghadapi kekhawatiran terkait akurasi prediksi cuaca yang mereka berikan kepada pelanggan mereka. Sebagai bagian dari tugasnya, perusahaan memberikan Doni sebuah dataset berisi informasi dan atribut-atribut terkait kondisi cuaca yang perlu diobservasi oleh-nya. Doni diminta untuk melakukan analisis statistika terhadap dataset tersebut guna membantu perusahaan dalam memahami keakuratan prediksi cuaca mereka serta membantu dalam melakukan berbagai pengujian hipotesis terkait faktor-faktor cuaca.

Atribut: Temperature, Humidity, Precipitation, Wind_Speed, Cloud_Coverage, Weather_Type, Wind_Direction, Pressure, UV_Index, Air_Quality, Visibility

Gunakan $\alpha = 0.05$

5. Hipotesis 1 sampel

- Perusahaan menerima beberapa keluhan bahwa prediksi terkait nilai humidity di suatu daerah seringkali tidak tepat. Hal tersebut berakibat pada kurangnya persiapan masyarakat dalam melakukan penyesuaian kondisi termal tertentu. Dapatkah Anda mengecek apakah rata-rata nilai Humidity lebih dari 75?
- Perusahaan mengeluarkan nilai rata-rata UV_index sebesar 3. Akan tetapi, mayoritas pelanggan mengeluhkan kulitnya terasa terbakar. Sebagai karyawan yang baik, periksalah apakah rata-rata UV_Index yang diamati memang tidak sama dengan 3 (sesuai laporan pelanggan)?
- Pemerintah setempat menyarankan agar penduduk menghindari aktivitas di luar ruangan di 5 hari terakhir (asumsi: data terbaru berada pada urutan paling awal) karena nilai rata-rata pressure diprediksi berada di angka 950. Periksa apakah nilai rata-rata pressure ?
- Apakah proporsi nilai Cloud_Coverage yang kurang dari 60 tidak sama dengan 35% dari total dataset, sesuai dengan himbauan yang diberikan oleh pemerintah setempat?

6. Hipotesis 2 sampel

Perusahaan ingin membandingkan kondisi cuaca di dua area geografis yang berbeda. Hal ini dapat dilakukan dengan membagi 1 dataset menjadi 2 bagian yang sama panjang.

- Dapatkah Anda memeriksa apakah rata-rata Humidity di Area A sama dengan rata-rata Humidity di Area B?

- Bandingkan rata-rata Wind Speed antara Area A dan Area B. Apakah rata-rata Wind Speed di Area A lebih tinggi daripada di Area B sebesar 5 mm?
- Perusahaan ingin membandingkan kualitas udara (Air_Quality) antara Area A dan Area B. Dapatkah Anda memeriksa apakah variansi Air_Quality di Area A sama dengan di Area B?
- Periksa apakah proporsi nilai precipitation yang kurang dari 7 pada daerah A lebih besar daripada kuantitas proporsi di daerah B dengan nilai yang sama?

Dataset Health

Mira adalah seorang peneliti kesehatan yang bekerja di sebuah lembaga riset medis yang terkemuka. Sebagai bagian dari tugasnya, Mira memiliki akses ke sebuah dataset yang berisi informasi tentang profil kesehatan dan gaya hidup dari sejumlah individu. Mira bertanggung jawab untuk melakukan analisis statistika terhadap dataset ini guna mendapatkan pemahaman yang lebih baik tentang faktor-faktor yang mempengaruhi kesehatan dan kualitas hidup individu. Selain itu, Mira juga diminta untuk mengidentifikasi pola dan hubungan yang signifikan antara variabel-variabel tersebut, serta untuk menjawab berbagai pertanyaan penelitian yang diajukan oleh lembaga riset.

Atribut: Age, Income, Gender, Education, Stress_Level, Exercise_Hours_Per_Week, Cholesterol_Level, Weight, Height, Blood_Pressure, Health_Status

Gunakan $\alpha = 0.05$

5. Hipotesis 1 sampel

- Lembaga riset saat ini sedang mempertanyakan data berat badan individu yang disimpan untuk kepentingan riset lanjutan. Identifikasilah apakah rata-rata berat badan pasien diatas 65 kg?
- Tekanan darah sistole yang normal berada pada rentang 120 mmHg. Lembaga riset perlu untuk memastikan apakah data individu yang diukur cukup normal. Periksalah apakah rata-rata tekanan darah sistole bernilai 120 mmHg?
- Periksalah apakah data 200 individu pertama pengujian (baris teratas) memiliki rata-rata waktu olahraga per minggu tidak sama dengan 15 jam?
- Apakah penduduk dengan pendapatan yang lebih besar dari Rp 7.500.000,00 tidak sama dengan 30% dari data keseluruhan individu?

6. Hipotesis 2 sampel

Lembaga riset membagi data individu menjadi dua bagian, yaitu data individu yang lebih awal masuk data penelitian (bagian atas) dan yang baru saja (bagian bawah).

- Periksa apakah rata-rata berat badan individu yang lebih awal masuk data penelitian sama dengan rata-rata berat badan individu yang masuk baru saja?
- Bagaimana dengan pendapatan individu, apakah pendapatan sistole individu yang lebih awal masuk data penelitian lebih besar Rp 1.250.000,00 dari yang baru saja masuk?

- Lembaga riset ingin membandingkan kondisi kesehatan individu dari dua bagian data. Apakah variansi tekanan darah individu yang lebih awal masuk data penelitian sama dengan yang baru saja masuk?
- Apakah proporsi waktu olahraga yang lebih dari 8 jam per minggu pada data individu awal lebih besar daripada kuantitas proporsi pada data individu akhir dengan waktu olahraga yang sama?

Dataset Phone

Markuis Graylee adalah seorang metuber yang membuat konten mengenai produk-produk elektronik. Untuk salah satu ide kontennya, ia ingin mereview *smartphone* yang telah dirilis oleh beberapa perusahaan. Sebagai pendukung penelitian kontennya, ia memiliki akses pada sebuah dataset yang berisi informasi-informasi relevan yang dapat diteliti untuk menilai produk-produk yang dirilis perusahaan. Markuis juga ingin menemukan pola-pola atau hubungan yang dimiliki oleh produk-produk tersebut agar ia dapat mengetahui tren yang ada pada dunia per-*smartphone*-an.

Atribut: *battery_power*, *clock_speed*, *ram*, *n_cores*, *use_time*, *px_width*, *px_height*, *brand*, 5g, *grade*, *price*

Gunakan $\alpha = 0.05$

5. Hipotesis 1 sampel

- Testimoni dari pengguna banyak yang menyatakan bahwa kapasitas baterai yang digunakan kurang dari sewajarnya. Periksa apakah *battery_power* memiliki rata-rata di atas 1800?
- Standar RAM yang dimiliki oleh suatu *smartphone* sekarang adalah 8 GB. Periksalah apakah rata-rata *ram smartphone* pada dataset adalah 8 GB?
- Periksa apakah 250 data pertama pada dataset memiliki rata-rata kecepatan clocking (*clock_speed*) tidak sama dengan 1!
- Periksalah apakah data *smartphone* dengan merek “appa” yang memiliki rata-rata waktu penggunaan (*use_time*) lebih dari 8.5 tidak sama dengan 35% dari data keseluruhan?

6. Hipotesis 2 sampel

Markuis berasumsi setengah bagian pertama dataset adalah *smartphone* generasi sebelumnya dan setengah bagian terakhir adalah *smartphone* generasi sekarang.

- Periksa apakah rata-rata jumlah core (*n_cores*) *smartphone* generasi sebelumnya sama dengan jumlah core *smartphone* generasi sekarang?
- Bagaimana dengan harga *smartphone*, apakah harga *smartphone* generasi sekarang lebih mahal 100 dari generasi sebelumnya?
- Apakah variansi dari tinggi *smartphone* (*px_height*) sama pada kedua generasi?


- Apakah proporsi kapasitas baterai (*battery_power*) *smartphone* yang lebih dari 2030 pada *smartphone* generasi sebelumnya lebih besar daripada proporsi kapasitas baterai (*battery_power*) *smartphone* yang lebih dari 2030 pada *smartphone* generasi sekarang?

Komponen Penilaian:

- Nomor 1, 2, 3 dan 4 : Kelengkapan jawaban dan ketepatan nilai
- Nomor 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, yang mencakup kesamaan jawaban akibat plagiarisme ataupun penggunaan tools seperti ChatGPT dan lain-lain.
3. Segala pertanyaan dapat ditanyakan melalui pranala

 QnA Probabilitas & Statistika 2024