Soal Mentoring Probability

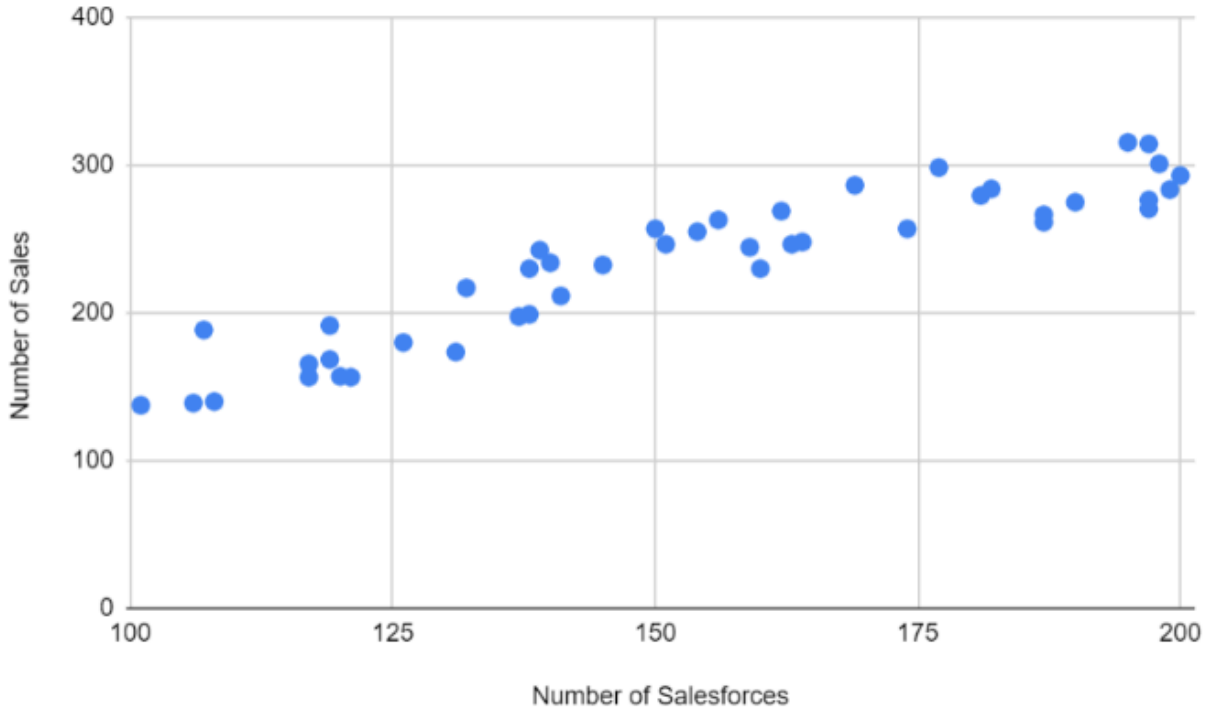
Job Preparation Program, Data Science, Pacmann AI

Catatan:

- Make a copy docs ini sebelum menjawab.
- Ada dua fokus dalam menjawab soal ini
    - Menemukan jawaban yang benar
    - Memahami cara mendapatkan jawaban yang benar
- Selama mengerjakan soal, Anda **wajib melampirkan** cara & proses menjawab. **BUKAN CUMA** pasang rumus, tapi **berikan elaborasi** kenapa Anda melakukan hal tersebut.
- Silahkan lampirkan proses Anda dengan menuliskan langsung pada docs atau melampirkan foto proses (yang dapat dibaca).
- Kumpulkan docs Anda sesuai dengan link submission yang tersedia.
- Selamat mengerjakan 🫡

---

1. [10 points] Setiap akhir periode penjualan, manager tim sales akan mencatat hasil kinerja tim sales. Beberapa diantaranya mencatat jumlah salesforce (salesman) yang turun dan jumlah sales yang dihasilkan. Grafik berikut menjelaskan hubungan antara variabel salesforce dan jumlah sales yang dihasilkan.



Mana pilihan yang benar dari dua opsi berikut dan sertakan alasannya!

a. Peningkatan jumlah salesforces menyebabkan peningkatan jumlah sales.
   Pernyataan ini masuk akal benar namun asumtif secara kausalitas. Jadi, pernyaaan ini tetap dianggap salah. Dapat diketahui dari grafik di atas bahwa secara visual, grafik menunjukkan tren naik. Ketika jumlah salesforces (X) meningkat, jumlah sales (Y) juga meningkat. Sehingga pernyataan ini benar bahwa peningkatan jumlah salesforces cenderung menyebabkan peningkatan jumlah sales.
   Namun, grafik ini hanya menunjukkan korelasi, bukan kausalitas langsung. Untuk bisa membuktikan bahwa peningkatan X “menyebabkan” peningkatan Y harus dibuktikan kausalitasnya melalui rancangan percobaan dengan controlled trial, difference-in-difference, dll.
b. Peningkatan jumlah sales menyebabkan peningkatan jumlah salesforces.
   Pernyataan ini salah. Ini merupakan kebalikan arah hubungan dari logika umum dalam manajemen tim sales. Biasanya, jumlah salesforces mempengaruhi jumlah sales, bukan sebaliknya.
   Dalam grafik ini, secara visual dan logika bisnis, jumlah salesforces adalah variabel X (independen), dan jumlah sales adalah variabel Y (dependen). Artinya, jumlah sales dipengaruhi oleh jumlah salesforces, bukan sebaliknya. Untuk membuktikan apakah benar “menyebabkan” diperlukan uji kausalitasnya melalui rancangan percobaan dengan controlled trial, difference-in-difference, dll.
c. Ada korelasi positif antara jumlah salesforces dan jumlah sales.

Pernyataan ini benar. Terdapat korelasi positif antara jumlah salesforces dan jumlah sales dapat dilihat dari bentuk kurva yang naik dari kiri ke kanan. Kurva menggambarkan kecenderungan ketika 1 variabel naik, maka variabel lainnya juga cenderung naik. Data menunjukkan hubungan positif yaitu semakin banyak jumlah salesforces, semakin banyak jumlah sales.

d. Ada korelasi negatif antara jumlah salesforces dan jumlah sales.
Pernyataan ini salah. Grafik menunjukkan semakin banyak jumlah salesforces, semakin banyak jumlah sales. Artinya korelasinya positif, bukan negatif. Korelasi negatif berarti ketika satu variabel naik, variabel lain turun, yang tidak terjadi di sini.

2. [15 points] Berapa covariance antara random variabel X ~ Uniform(0, 1) dan Y apabila diketahui Y = 2.X? Jelaskan cara Anda mencari jawabannya.
X ~ Uniform(0, 1)
Y = 2.X

Yang akan dihitung covariance antara X dan Y, yaitu Cov(X,Y)

1. Rumus dasar covariance
Cov(X,Y) = E[XY] - E[X] . E[Y]

Y = 2X
XY = X . 2X
= $2X^2$
Jadi,
E[XY] = E[$2X^2$]
= 2 . E[X^2]

2. Cari E[X] dan E[X^2] untuk X ~ Uniform(0, 1)
E[X] = (a+b)/2
= (0+1)/2
= ½
E[X^2] = $1/(b-a) \int_a^b x^2 dx = \int_0^1 x^2 dx$
= $[x^3/3]_0^1$
= 1/3
3. Hitung E[XY]
E[XY] = 2 . E[X^2]
= 2 . ⅓
= ⅔

4. Hitung E[Y]
Karena Y = 2X
E[Y] = E[2X]
= 2 . E[X]

= 2 . ½
= 1

5. Hitung covariance
   Cov(X,Y) = E[XY] - E[X] . E[Y]
   = ⅔ - ½ . 1
   = ⅔ - ½
   = (4-3)/6
   = ⅙

Jadi, covariance antara random variabel X ~ Uniform(0, 1) dan Y apabila diketahui Y = 2X adalah ⅙. Hal ini dapat diketahui juga dari rumus eksplisit. Karena Y =2X maka kita bisa menggunakan format Cov(X,aX) = a . Var(X). Kita dapat tahu bahwa Var(X) untuk uniform (0,1) adalah 1/12. Maka Cov (X,2X) adalah 2 x Var(X) = 2 x 1/12 = ⅙.

3. [30 points] Tim marketing biasa melakukan campaign dengan mengirimkan email yang berisi penjelasan produk ke calon pembeli. Tim marketing mengukur kesuksesan campaign dari metric conversion rate.

   conversion rate = jumlah email yang melakukan pembelian / jumlah total email yang dikirim

   Selama 5 tahun menjadi bagian dari tim marketing, Anda tahu kalau conversion rate rata-rata sekitar 11.2%. Namun Anda menyadari conversion rate bulan ini adalah 11.7% dari 10.000 email yang dikirim.

   Anda ingin tau apakah conversion rate-nya memang meningkat di bulan ini. Lakukanlah uji hipotesis dengan confidence level 95%!

   a. [10 points] Apa yang diuji dalam kasus ini?
      Yang diuji adalah apakah conversion rate bulan ini (11.7%) secara statistik meningkat (lebih tinggi) dari conversion rate rata-rata historis (11.2%). Kita ingin mengetahui apakah kenaikan ini signifikan atau hanya karena variasi acak dari pengambilan sampel.

   b. [10 points] Tulis hipotesis yang diuji (H_0 dan H_a) dalam kasus ini
      H_0: $p = p_0$
      H_0: $p = 11.2\%$
      Tidak ada peningkatan conversion rate bulan ini. Jadi, dari 11.2% menjadi 11.7% hanya karena variasi acak dari pengambilan sampel (status quo)
      H_a: $p > p_0$
      H_a: $p > 11.2\%$
      Ada peningkatan conversion rate bulan ini.

c. [10 points] Lakukan uji hipotesis dan rangkum hasilnya
Diketahui:
p^ = 0.117 (proporsi bulan ini)
p_0 = 0.112 (proporsi historis)
n = 10.000
α = 0.05

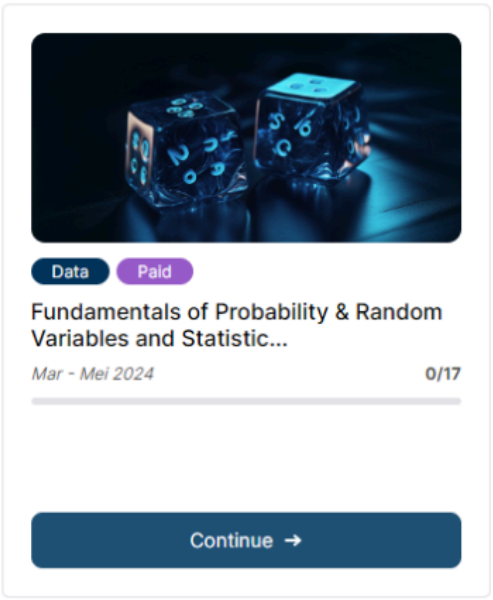
Maka, akan dilakukan uji proporsi Z karena kita akan menguji proporsi email yang berujung pada pembelian dengan asumsi p_0 diketahui dan ukuran sampel besar.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

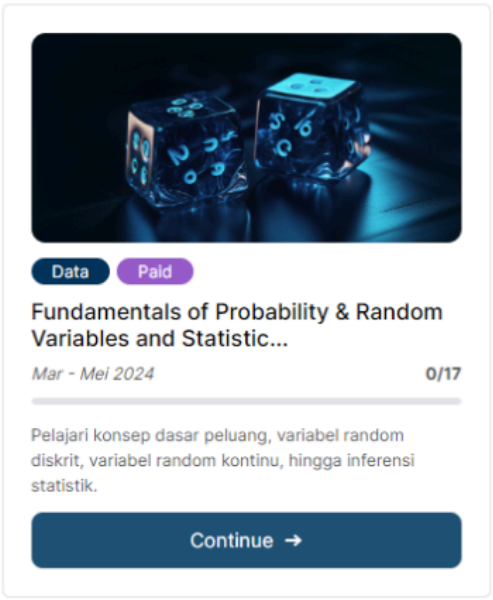
Z = (0.117 - 0.112) / √ (0.112(1-0.112))/10.000)
= 0.005 / √ 0.0000099456
= 0.005 / 0.003154
= 1.585
Lihat critical value untuk α = 0.05 yaitu critical value z = 1.645
Z = 1.585 < 1.645 maka gagal tolak H_0

4. [45 points] Pacmann merasa sukses apabila siswanya belajar dengan baik di platform belajar. Hal yang bisa Pacmann lakukan adalah meningkatkan jumlah siswa yang belajar di platform belajar. Anda (bagian dari tim product Pacmann) memutuskan untuk melakukan eksperimen dengan tujuan meningkatkan siswa yang belajar di platform.

   Anda terpikir untuk melakukan eksperimen sederhana yakni **menambahkan deskripsi** pada *card* kelas pada halaman utama platform belajar siswa. Anda melakukan eksperimen ini selama 1 bulan dan mengukur metrik *conversion rate*.



Card kelas **sebelum**
Tanpa deskripsi



Card kelas **sesudah**
Dengan deskripsi

Conversion rate = jumlah siswa yang mengakses platform dan klik tombol “continue” / jumlah siswa yang mengakses platform

Hasil eksperimen dapat Anda akses pada data berikut: [pacmann_lms_experiment.csv]

user_id	timestamp	group	converted
95161	2024-04-01 00:29:34	treatment	0
42631	2024-04-01 00:54:48	control	1
29853	2024-04-01 01:07:16	control	0
80425	2024-04-01 01:07:45	treatment	0
66743	2024-04-01 01:21:00	treatment	0

Dengan definisi data:

- **user_id** : user ID yang mengakses platform belajar
- **timestamp** : Waktu user dalam mengakses platform belajar atau klik tombol continue

- **group** : Grup card kelas yang ditampilkan. Apabila grup = “control” maka user melihat card **tanpa** deskripsi. Apabila grup = “treatment” maka user melihat card **dengan** deskripsi.
- **converted** : Status user. Apabila 1, artinya user klik tombol “continue” pada card kelas.

Anda ingin mengetahui apakah menambahkan deskripsi pada *card* kelas dapat meningkatkan metrik conversion rate. Lakukanlah uji hipotesis dengan confidence level 95%!
Hint: Ini adalah eksperimen A/B Testing

a. [10 points] Apa yang diuji dalam kasus ini?
   Yang diuji adalah apakah conversion rate platform dengan card deskripsi (grup treatment) secara statistik meningkat (lebih tinggi) dari conversion rate platform sebelum menggunakan deskripsi di card kelas (grup control). Kita ingin mengetahui apakah kenaikan ini signifikan atau hanya karena variasi acak dari pengambilan sampel. Ini adalah uji proporsi dua sampel. Sampel siswa yang klik “continue” karena card dengan deskripsi dan sampel siswa yang klik “continue” sebelum card dideskripsikan.

b. [10 points] Tulis hipotesis yang diuji (H_0 dan H_a) dalam kasus ini
   $p_1 = p_c$ = conversion rate grup control
   $p_2 = p_t$ = conversion rate grup treatment

   H_0: $p_t <= p_c$
   Penambahan deskripsi di card kelas tidak meningkatkan conversion rate (status quo)
   H_a: $p_t > p_c$
   Ada peningkatan conversion rate karena deskripsi di card kelas.

c. [25 points] Lakukan uji hipotesis dan rangkum hasilnya
   Akan dilakukan uji proporsi Z dua sampel karena kita akan membandingkan proporsi dari 2 grup yaitu grup control dan treatment dengan sampel besar.

   1. Menyatakan hipotesis
      H_0: $p_t <= p_c$
      Penambahan deskripsi di card kelas tidak meningkatkan conversion rate (status quo)
      H_a: $p_t > p_c$
      Ada peningkatan conversion rate karena deskripsi di card kelas.

   2. Olah dataset dengan python

```
import pandas as pd
import numpy as np
from scipy.stats import norm

# 1. Baca data
from google.colab import drive
drive.mount('/content/drive')

file_path = '/content/drive//My Drive/PACMANN/PORTOFOLIO/PROBABILITY/pacmann_lms_experiment.csv'
df = pd.read_csv(file_path)

# Lihat 5 bar  is pertama
df.head()
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/
```

	user_id	timestamp	group	converted
0	95161	2024-04-01 00:29:34	treatment	0
1	42631	2024-04-01 00:54:48	control	1
2	29853	2024-04-01 01:07:16	control	0
3	80425	2024-04-01 01:07:45	treatment	0
4	66743	2024-04-01 01:21:00	treatment	0

3. Hitung statistik deskriptif per grup untuk uji statistik

```
[6]  # 2. Hitung statistik per grup
     group_stats = df.groupby("group")["converted"].agg(
         total_users='count',
         total_converted='sum',
         conversion_rate='mean'
     ).reset_index()

     print("Ringkasan per grup:")
     print(group_stats)
```

```
Ringkasan per grup:
       group  total_users  total_converted  conversion_rate
0    control         1768              236         0.133484
1  treatment         1767              260         0.147142
```

x_c = 236
x_t = 260
n_c = 1768
n_t = 1767

4. Hitung proporsi keseluruhan keberhasilan

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

p^ = (x_c + $x_{t)}$ / (n_c + $n_{t)}$
= 236 + 260 / 1768 + 1767
= 496 / 3535
= 0.140

5. Hitung uji statistik

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$p^_c$ = x_c / n_c
= 236 / 1768
= 0.133

$p^_t$ = x_t / n_t
= 260 / 1767
= 0.147

Z = $p^_t$ - $p^_c$ / √ p^ (1 - p^) ($1/n_t$ + $1/n_c$)
= 0.147 - 0.133 / √ (0.140)(1 - 0.140) (1/1767 + 1/1768)
= 0.014 / √ (0.140) (0.86) (0.00057 + 0.00057)
= 0.014 / √ (0.140) (0.86) (0.00114)
= 0.014 / √ 0.000137
= 0.014 / 0.01172
= 1.1945

6. Menentukan aturan keputusan
Aturan keputusan (taraf signifikansi 5%)
Z-crit = 1.645 untuk one-tailed test uji pihak kanan
Tolak jika Z > 1.645
Karena Z = 1.1945 < 1.645
Maka gagal menolak H_0

7. Menyatakan rejection decision
Gagal Tolak H_0 di taraf signifikansi 5% karena Z < 1.645 (Z = 1.1945)

8. Kesimpulan

Tidak terdapat bukti yang cukup kuat secara statistik bahwa menambahkan deskripsi pada card kelas secara signifikan meningkatkan conversion rate siswa.