

Statistika inferensi

- Statistika deskriptif: mengidentifikasi hubungan dari suatu data, menarik kesimpulan dari sampel.
- Statistika inferensi: menarik kesimpulan tentang populasi.

Nilai p

Nilai p merupakan peluang:

- Jika hipotesis awal benar, maka sampel dapat menghasilkan estimasi.
- Menerangkan kemungkinan kita mendapatkan hasil.

Contoh:

Ada warung Zominos Pizza:

- **Komplain** pelanggan bahwa keju nya tidak cukup banyak.
- **Regulasi**: Harus pakai 100 gr keju untuk setiap pizza.

Namun, manajer Zominos tidak dapat mengunjungi seluruh warung di kota tersebut, sehingga diambilah sampel dari beberapa warung.

Pada kasus ini:

- H_0 : (*Trying to provide evidence against*) Menolak klaim pelanggan (Cukup kok keju nya).
- H_1 : Apa yang hendak kita buktikan, yakni tidak cukup keju untuk setiap Pizza.

Significance level (α) = 0,05

Jika nilai p dibawah α , maka kita dapat menolak hipotesis awal.

```
library(dplyr)
df <- read.csv("../data/ZominosCheese.csv")
head(df)
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':
```

```
filter, lag
```

```
The following objects are masked from 'package:base':
```

```
intersect, setdiff, setequal, union
```

GRAMSCHEESE

85

100

85

98

94

97

```
# peraturannya harus pakai keju 100 gr (mu)  
t.test(df,mu=100)
```

One Sample t-test

```
data: df  
t = -0.96396, df = 29, p-value = 0.343  
alternative hypothesis: true mean is not equal to 100  
95 percent confidence interval:  
 95.94179 101.45821  
sample estimates:  
mean of x  
 98.7
```

Nilai p -nya 0,0343, nilai rata2nya = 98,7. Nilai $p > 0,05$, maka kita menolak hipotesis awal.

Derajat kebebasan

Derajat kebebasan merujuk pada ukuran maksimum dari nilai - nilai logikal yang bersifat independen.

$$df = n - 1$$

```
data <- read.csv("../data/ZominosCheese.csv")
head(data)
t.test(data, mu=100)
```

GRAMSCHEESE

85

100

85

98

94

97

One Sample t-test

```
data: data
t = -0.96396, df = 29, p-value = 0.343
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 95.94179 101.45821
sample estimates:
mean of x
 98.7
```

$$df = 29$$

```
df = length(data$GramsCheese) - 1
df
```

29

Selang dan level kepercayaan

- Selang kepercayaan mengukur derajat ketidakpastian atau kepastian.
- Level kepercayaan merupakan persentase dari peluang atau kepastian.

```
t.test(data, mu=100)
```

One Sample t-test

```
data: data
t = -0.96396, df = 29, p-value = 0.343
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 95.94179 101.45821
sample estimates:
mean of x
 98.7
```

Selang kepercayaan: 95,94179 hingga 101,45821

Pengujian hipotesis

Pengujian hipotesis merupakan prosedur di dalam statistika inferensi

- H_0 (hipotesis awal): Dua populasi tidak berbeda jika merujuk pada sifat tertentu.
- H_1 (hipotesis alternatif): Efek tertentu terjadi pada dua buah populasi.

Jika nilai $-p$ dibawah *significance level*, maka kita dapat menolak hipotesis awal.

Untuk melakukan pengujian hipotesis, kita menggunakan uji $-t$

Uji $-t$ adalah salah satu teknik statistika inferensi yang digunakan untuk menentukan perbedaan signifikan rata - rata antar dua kelompok.

Terdapat 3 jenis uji $-t$:

- Uji sampel tunggal.
- Uji dua sampel.
- Uji sampel berpasangan.

Studi kasus

Zominos mempunyai penawaran beli 1 gratis 1 pada hari tertentu untuk meningkatkan penjualan.

- Sampel 1: Penjualan pada hari - hari promosi.
- Sampel 2: Penjualan pada hari - hari biasa.

Berikut adalah hipotesis nya:

- H_0 : Tidak ada perbedaan antara penjualan di hari - hari promosi dan hari - hari biasa (Penjualan promo = Penjualan biasa).
- H_1 : Terdapat perbedaan (Penjualan promo \neq Penjualan biasa).

(Disebut sebagai *two-tail test*).

- H_0 : Penjualan promo - Penjualan biasa = 0
- H_1 : Penjualan promo - Penjualan biasa \neq 0

Jika kita tidak tertarik pada penjualan yang berkurang pada hari - hari promosi, maka uji hipotesis dapat diubah formulasinya:

- H_0 : Penjualan promo - Penjualan biasa \leq 0.
- H_1 : Penjualan promo - Penjualan biasa $>$ 0.

```
df <- read.csv("../data/ZominosSales.csv")
head(df)
```

OFFERDAYS	NONOFFERDAYS
248.3	215.1
335.2	300.0
338.0	320.6
285.3	276.6
322.2	282.9
283.6	288.1

```
t.test(df$OfferDays, df$NonOfferDays) # Two-tail test
```

Welch Two Sample t-test

```
data: df$OfferDays and df$NonOfferDays
t = 2.6105, df = 17.229, p-value = 0.01814
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 7.997416 75.042584
sample estimates:
mean of x mean of y
 307.35    265.83
```

Nilai $-p < 0,05$, maka kita menolak H_0 (Penjualan promo - Penjualan biasa = 0).

```
# One-tail test
t.test(df$OfferDays, mu = 265.83) # mu = NonOfferDays
```

One Sample t-test

```
data: df$OfferDays
t = 4.1576, df = 9, p-value = 0.002456
alternative hypothesis: true mean is not equal to 265.83
95 percent confidence interval:
 284.7592 329.9408
sample estimates:
mean of x
 307.35
```

Nilai $-p < 0,05$, maka kita menolak H_0 (Penjualan promo - Penjualan biasa \leq 0).

Uji chi-kuadrat

Uji chi-kuadrat menilai apakah baris dan kolom pada tabel kontingensi berhubungan secara signifikan secara statistik.

Terdapat dua jenis uji chi-kuadrat:

- *Test of independence*
- *Goodness-of-Fit Test*
- H_0 : Variabel baris dan kolom dari tabel kontingensi bersifat independen.
- H_1 : Variabel baris dan kolom saling bergantung.

Jika nilai uji chi-kuadrat lebih besar dari *significance level*, maka mengindikasikan bahwa kolom dan baris saling berhubungan satu dengan yang lain.

```
ratings <-
  factor(c(2,4,3,3,2,1,1,2,3,4,2,3,3,4,1,3,2,1,4,3,2,4))
ratings
```

1. 2
2. 4
3. 3
4. 3
5. 2

6. 1
7. 1
8. 2
9. 3
10. 4
11. 2
12. 3
13. 3
14. 4
15. 1
16. 3
17. 2
18. 1
19. 4
20. 3
21. 2
22. 4

► **Levels:**

```
kursus <-  
factor(c(1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,1,1,1,1,1,0))  
kursus
```

1. 1
2. 1
3. 1
4. 0
5. 0
6. 0
7. 0
8. 0
9. 1
10. 1
11. 1
12. 0
13. 0
14. 0
15. 0
16. 0
17. 1
18. 1
19. 1
20. 1
21. 1
22. 0

► **Levels:**

```
levels(kursus) <- c('R', 'Python')
```

```
kursus
```

1. Python
2. Python
3. Python
4. R
5. R
6. R
7. R
8. R
9. Python
10. Python
11. Python
12. R
13. R
14. R
15. R
16. R
17. Python
18. Python
19. Python
20. Python
21. Python
22. R

► **Levels:**

```
data <- table(ratings, kursus)
data
```

	kursus	
ratings	R	Python
1	3	1
2	2	4
3	4	3
4	2	3

```
chisq.test(data)
```

```
Warning message in chisq.test(data):
"Chi-squared approximation may be incorrect"
```

Pearson's Chi-squared test


```
data: data
X-squared = 2.0095, df = 3, p-value = 0.5704
```

Nilai- p : 0,5704. Maka dapat disimpulkan jika variabel baris dan kolom tidak independen. Secara implisit menyatakan ada hubungan antara kursus dengan ratings.

```
chisq.test(data, simulate.p.value=T)
# Karena datanya kurang, kita simulasikan untuk jumlah
data replikasi yang lebih besar
```

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

```
data: data
X-squared = 2.0095, df = NA, p-value = 0.7076
```