STK351
Pengantar Analisis Data Kategorik

Farit Mochamad Afendi
08128592194 – fmafendi@apps.ipb.ac.id

Deskripsi MK

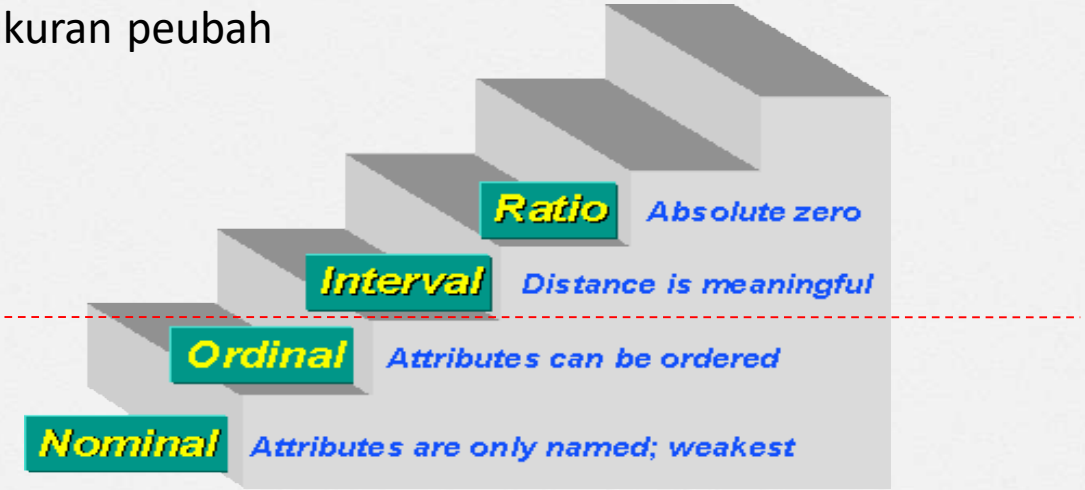
- Mata kuliah ini membahas tentang metode statistika untuk data kategorik yang mencakup metode yang memiliki peran penting dalam perjalanan sejarah statistika seperti uji Khi-kuadrat sampai ke model analisis statistika yang berkembang sejalan perkembangan mutakhir dari teknologi komputasi seperti model regresi logistik

Mengapa analisis data kategorik?

Skala pengukuran peubah

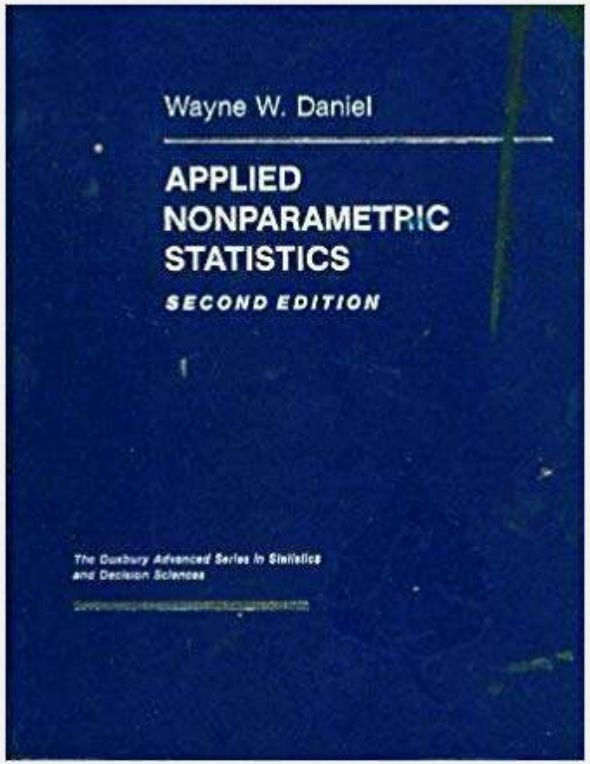
Numerik

Kategorik

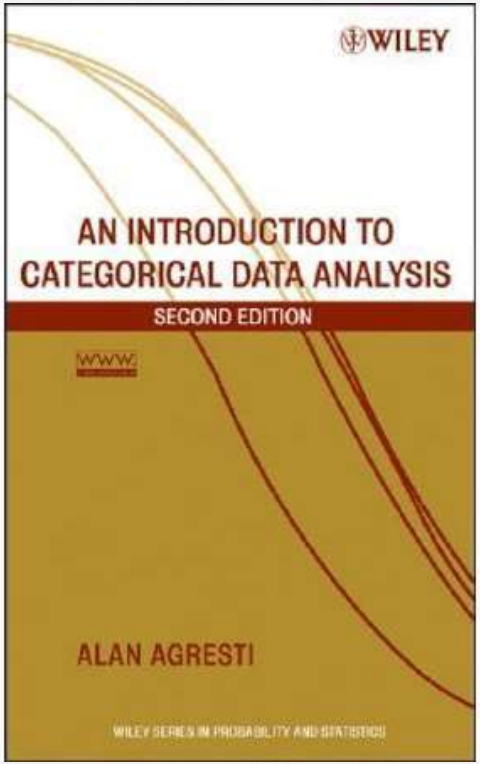


Buku referensi

1



2



No.	Pokok Bahasan	Sub Pokok Bahasan	Perkiraan Waktu (menit)	Daftar Kepustakaan
1.	Pendahuluan Statistika Nonparametrik	▪ Apa dan mengapa Statistika Non-parametrik ▪ Keterkaitan non-parametrik dengan analisis data kategorik ▪ Uji mengenai nilai-tengah: pembandingan metode parametrik dan nonparametrik	1 x (2 x 50')	1: Bab 1 – 2
2.	Prosedur uji nonparametrik untuk pembandingan nilai-tengah dua populasi	▪ Prosedur yang melibatkan dua contoh bebas ▪ Prosedur yang melibatkan dua contoh berpasangan ▪ Korelasi Spearman	1 x (2 x 50')	1: Bab 3 – 4, 9
3.	Statistik Khi-kuadrat	▪ Uji Khi-kuadrat untuk tabel frekuensi (sebaran seragam, sebaran binomial, sebaran Poisson) ▪ Uji Khi-kuadrat untuk kebebasan dan kehomogenan	1 x (2 x 50')	1: Bab 5,8

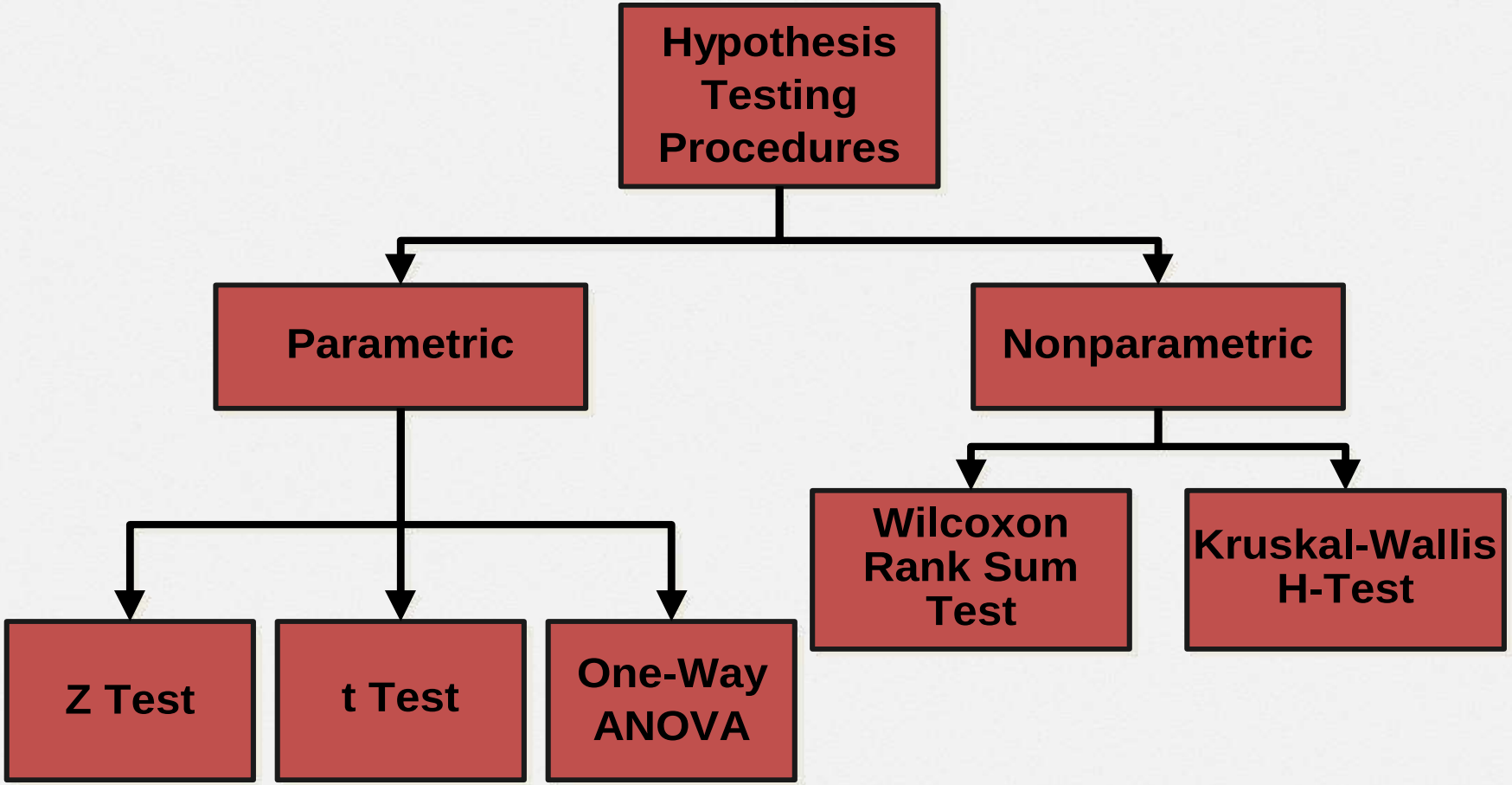
No.	Pokok Bahasan	Sub Pokok Bahasan	Perkiraan Waktu (menit)	Daftar Kepustakaan
4.	Data Respon Kategorik	▪ Apa dan mengapa analisis data kategorik ▪ Peubah Respon dan Peubah Penjelas ▪ Skala Nominal dan Skala Ordinal ▪ Review Sebaran Binomial dan Sebaran Multinomial	1 x (2 x 50')	2: Bab 1.1-1.2
5.	Inferensi untuk Parameter Proporsi	▪ Fungsi Kemungkinan (likelihood function) ▪ Uji Statistik untuk Parameter Binomial ▪ Selang Kepercayaan untuk Parameter Binomial ▪ Inferensi untuk Ukuran Contoh Kecil	1 x (2 x 50')	2: Bab 1.3, 1.4.3

No.	Pokok Bahasan	Sub Pokok Bahasan	Perkiraan Waktu (menit)	Daftar Kepustakaan
6.	Tabel Kontingensi 2x2	▪ Peluang Bersama, Peluang Marjinal, dan Peluang Bersyarat ▪ Kepekaan dan Kekhususan dalam Uji Diagnostik ▪ Kebebasan	1 x (2 x 50')	2: Bab 2.1
7.	Tabel Kontingensi 2x2	▪ Percontohan Binomial dan Multinomial ▪ Beda Proporsi ▪ Risiko Relatif	1 x (2 x 50')	2: Bab 2.1.5 Bab 2.2
8.	Tabel Kontingensi 2x2	▪ Rasio Odd ▪ Uji Kebebasan Khi-kuadrat	1 x (2 x 50')	2: Bab 2.3, 2.4
9.	Tabel Kontingensi 2x2	▪ Uji Kebebasan untuk Data Ordinal ▪ Uji Eksak untuk Ukuran Contoh Kecil	1 x (2 x 50')	2: Bab 2.5, 2.6

No.	Pokok Bahasan	Sub Pokok Bahasan	Perkiraan Waktu (menit)	Daftar Kepustakaan
10.	Regresi Logistik	▪ Interpretasi Model Regresi Logistik ▪ Inferensi untuk Regresi Logistik	1 x (2 x 50')	2: Bab 4.1, 4.2
11.	Regresi Logistik	▪ Prediktor Kategorik ▪ Uji Cochran-Mantel Haenszel ▪ Uji Kehomogenan Rasio Odd	1 x (2 x 50')	2: Bab 4.3
12.	Regresi Logistik Berganda	▪ Contoh Regresi Logistik Ganda ▪ Pembandingan Model	1 x (2 x 50')	2: Bab 4.4.1, 4.4.2
13.	Regresi Logistik Berganda	▪ Prediktor Kuantitatif dalam Regresi Logistik ▪ Model dengan Interaksi	1 x (2 x 50')	2: Bab 4.4.3, 4.4.4
14.	Penerapan Model Regresi Logistik	▪ Strategi Pemilihan Model ▪ Pemeriksaan Kecocokan Model	1 x (2 x 50')	2: Bab 5.1, 5.2

Pengajar

- Farit Mochamad Afendi
- Asep Saefuddin
- Pika Silvianti


Hypothesis Testing Procedures
Parametric
Nonparametric
Z Test
t Test
One-Way ANOVA
Wilcoxon Rank Sum Test
Kruskal-Wallis H-Test

Parametric Test Procedures

1. Involve Population Parameters (Mean)

2. Have Stringent Assumptions (Normality)

3. Examples: Z Test, t Test, χ^2 Test, F test

Nonparametric Test Procedures

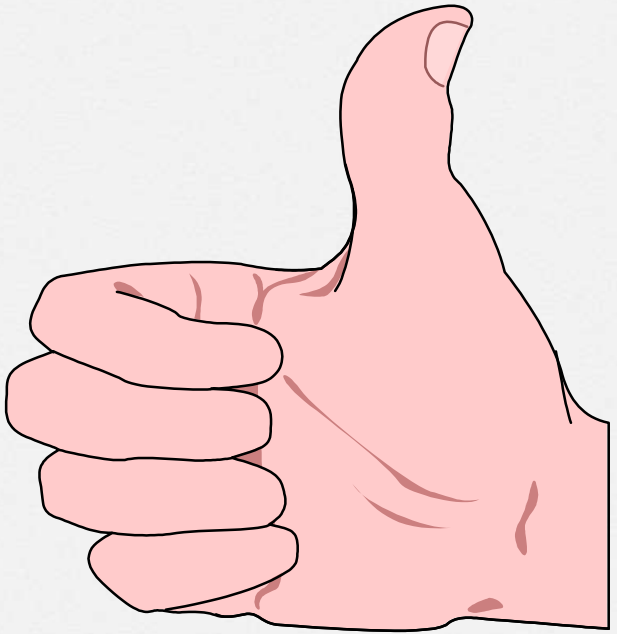
1. Do Not Involve Population Parameters
   Example: Probability Distributions, Independence

2. Data Measured on Any Scale (Ratio or Interval, Ordinal or Nominal)

3. Example: Wilcoxon Rank Sum Test

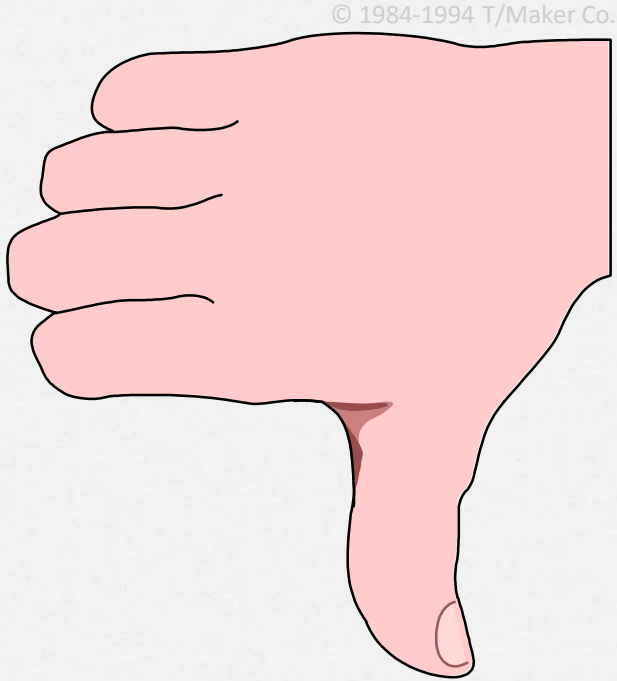
Advantages of Nonparametric Tests

1. Used With All Scales
2. Easier to Compute
3. Make Fewer Assumptions
4. Need Not Involve Population Parameters
5. Results May Be as Exact as Parametric Procedures



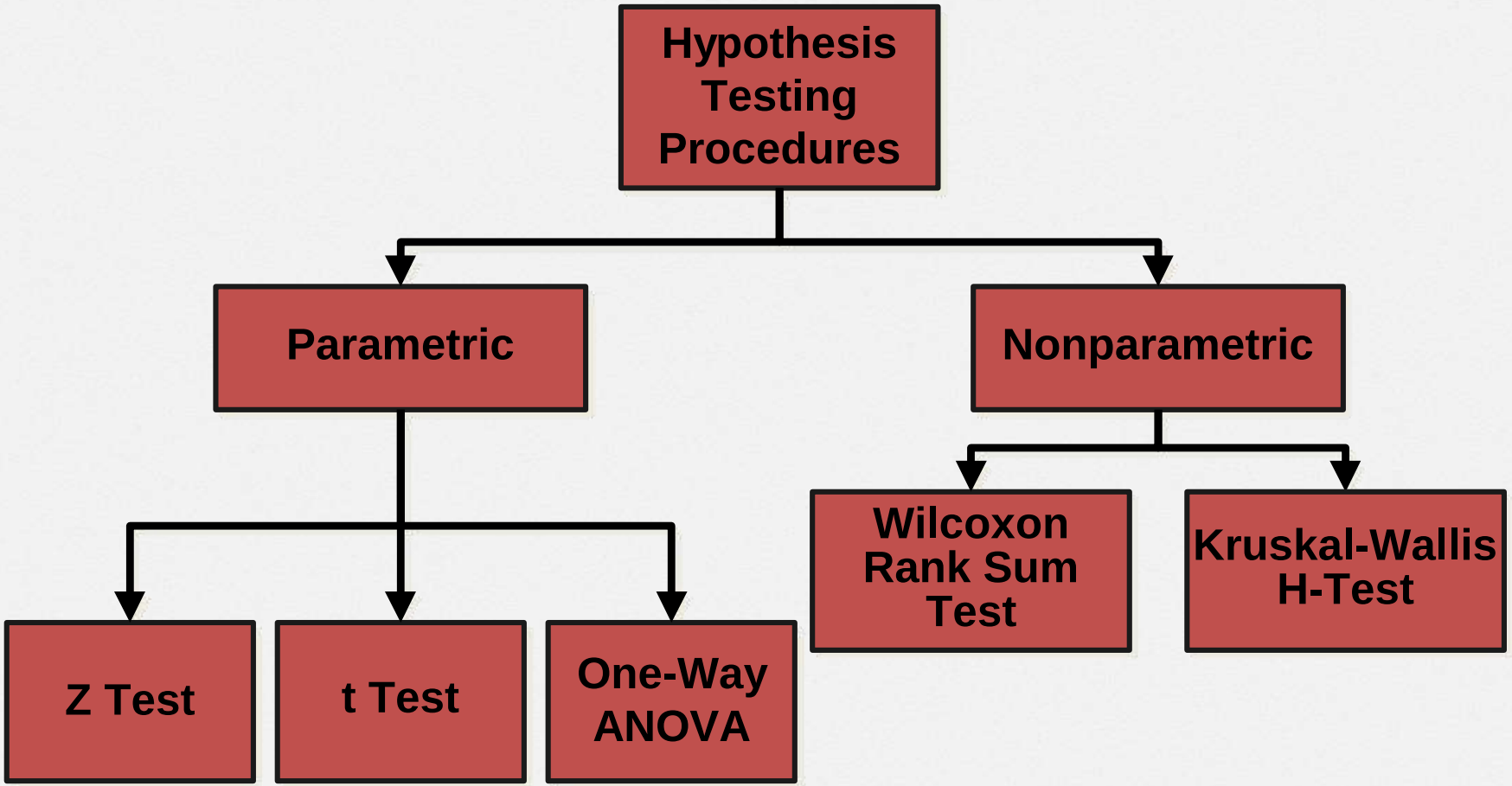
Disadvantages of Nonparametric Tests

1. May Waste Information
   Parametric model more efficient if data Permit
2. Difficult to Compute by hand for Large Samples
3. Tables Not Widely Available



Uji Nonparametrik

Farit Mochamad Afendi
08128592194 – fmafendi@apps.ipb.ac.id


Hypothesis Testing Procedures
Parametric
Nonparametric
Z Test
t Test
One-Way ANOVA
Wilcoxon Rank Sum Test
Kruskal-Wallis H-Test

Parametric Test Procedures

1. Involve Population Parameters (Mean)

2. Have Stringent Assumptions (Normality)

3. Examples: Z Test, t Test, χ^2 Test, F test

Nonparametric Test Procedures

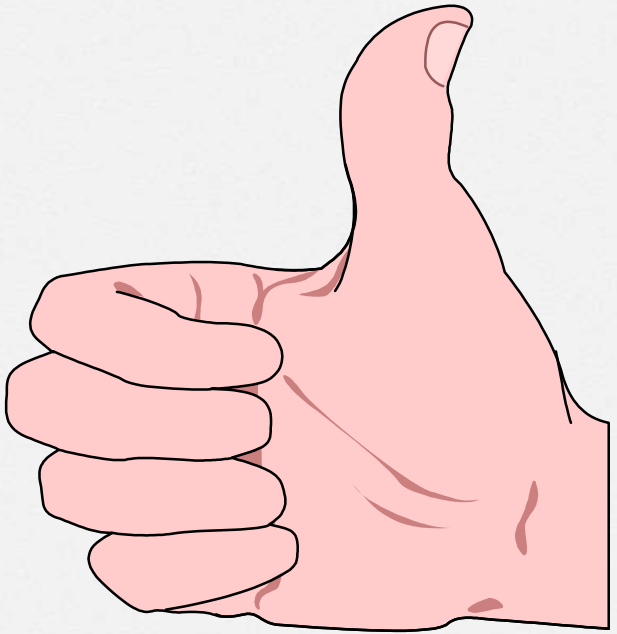
1. Do Not Involve Population Parameters
   Example: Probability Distributions, Independence

2. Data Measured on Any Scale (Ratio or Interval, Ordinal or Nominal)

3. Example: Wilcoxon Rank Sum Test

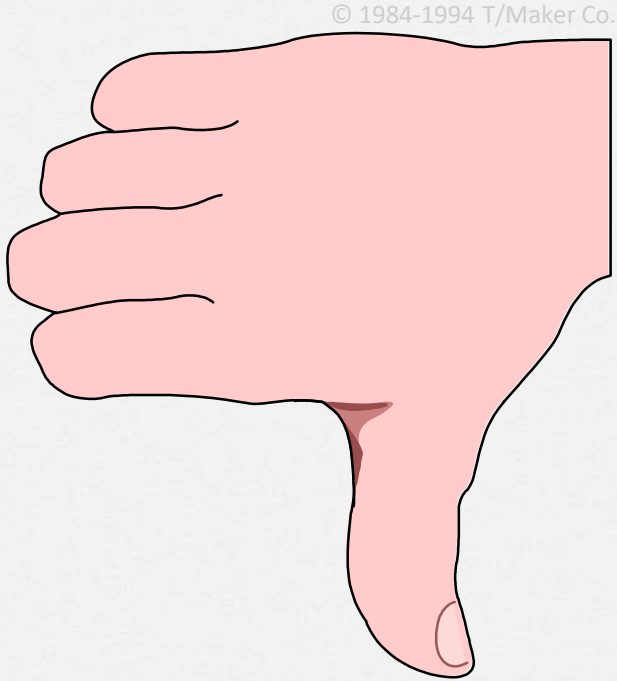
Advantages of Nonparametric Tests

1. Used With All Scales
2. Easier to Compute
3. Make Fewer Assumptions
4. Need Not Involve Population Parameters
5. Results May Be as Exact as Parametric Procedures



Disadvantages of Nonparametric Tests

1. May Waste Information
   Parametric model more efficient if data Permit
2. Difficult to Compute by hand for Large Samples
3. Tables Not Widely Available



UJI TANDA

Uji Tanda

- Berkaitan dengan pengujian nilai tengah:
  - satu populasi
  - dua populasi dengan teknik percontohan berpasangan
- Sebagaimana uji non parametrik untuk nilai tengah lainnya, fokus uji ini adalah median populasi
- Dinamakan uji tanda karena prosedur uji ini menandai satu persatu amatan (Xi) sesuai perbandingannya dengan median yang diujikan (M):
  - tanda + bila Xi > M
  - tanda – bila Xi < M
- Amatan yang nilainya persis sama dengan M, tidak disertakan dalam analisis dan mengurangi ukuran contoh efektif

Hipotesis yang diuji

A. $H_0: M \leq M_0$ vs $H_1: M > M_0$
B. $H_0: M \geq M_0$ vs $H_1: M < M_0$
C. $H_0: M = M_0$ vs $H_1: M \neq M_0$

Ilustrasi 1

- Benarkah nilai tengah omset bulanan UKM Kota Bogor sebesar Rp 10 juta?
- 10 contoh UKM dipilih untuk verifikasi dengan omset bulanannya: 8.1, 7.8, 13.5, 12.1, 10.0, 5.3, 6.9, 9.5, 11.5, 10.1
- H_0: M = 10 vs H_1: M ≠ 10

Ilustrasi 1

8.1	–	
7.8	–	
13.5	+	
12.1	+	$n_- = 5$
10.0	X	$n_+ = 4$
5.3	–	$n_{eff} = 9$
6.9	–	
9.5	–	
11.5	+	
10.1	+	

- Bila median populasi benar sebesar 10, maka n- dengan n+ mestinya relatif berimbang
- Ketidakberimbangan keduanya mengindikasikan H1 lebih mungkin benar
- Nilai-p: peluang mendapati komposisi seperti yang didapatkan dari amatan atau yang lebih tidak berimbang lagi dari itu
- Nilai-p = $2P(X \leq 4)$
  - $X \sim$ Binom(9, 0.5)
  - 4 karena n+ < n-.
  - Dikalikan dua karena dua arah
- Nilai-p = 2 (0.5) = 1
- Bila taraf nyata α = 5%, maka H0 tidak ditolak → **belum cukup bukti** untuk menyatakan nilai tengah omset bulanan UKM Kota Bogor **tidak** sebesar Rp 10 juta

Ilustrasi 2

- Program pendampingan dilakukan pada UKM Kota Bogor agar terjadi peningkatan pendapatan yang mereka terima
- Untuk keperluan ini, 10 contoh UKM diamati omsetnya sebelum dan sesudah program pendampingan.

Ilustrasi 2

Sebelum [1]	Sesudah [2]	Selisih ([2]-[1])
8.1	8.2	0.1
7.8	7.7	-0.1
13.5	13.5	0.0
12.1	14.0	1.9
10.0	12.0	2.0
5.3	5.7	0.4
6.9	6.5	-0.4
9.5	9.2	-0.3
11.5	12.0	0.5
10.1	11.0	0.9

- Terjadi peningkatan pendapatan berarti selisih > 0
- Hipotesis yang diuji:

  H_0: $M \leq 0$ (tidak terjadi peningkatan omset)

  H_1: $M > 0$ (terjadi peningkatan omset)

Ilustrasi 2

Selisih	Tanda
0.1	+
-0.1	–
0.0	X
1.9	+
2.0	+
0.4	+
-0.4	–
-0.3	–
0.5	+
0.9	+

$n_- = 3$
$n_+ = 6$
$n_{eff} = 9$

- Bila benar tidak terjadi peningkatan omset, maka n+ mestinya relatif sedikit
- Nilai n+ mengindikasikan seberapa mungkin H1 benar
- Nilai-p: peluang mendapati n+ seperti yang didapatkan dari amatan atau yang lebih besar lagi dari itu
- Nilai-p = $P(X \geq 6)$
  - $X \sim$ Binom(9, 0.5)
  - 6 karena n+
- Nilai-p = 0.2539
- Bila taraf nyata α = 5%, maka H0 tidak ditolak → **belum cukup bukti** untuk menyatakan **terjadi peningkatan** omset bulanan UKM Kota Bogor sesudah program pendampingan tersebut

UJI PERINGKAT BERTANDA WILCOXON

Ide Uji Peringkat Bertanda Wilcoxon

- Uji tanda hanya menandai tiap amatan sesuai nilai relatifnya terhadap nilai median yang diujikan, namun mengabaikan besarnya selisih keduanya
- Selayaknya, nilai selisih ini diperhitungkan dalam pengujian karena ikut berkontribusi terhadap perbedaan nilai amatan dengan median yang diuji
- Uji ini berupaya menampung perbedaan ini lewat peringkat dari nilai amatan

Ilustrasi

X_i	$X_i - M_0$	$\lvert X_i - M_0 \rvert$	tanda
8.1	-1.9	1.9	–
7.8	-2.2	2.2	–
13.5	+3.5	3.5	+
12.1	+2.1	2.1	+
10.0	0	0	X
5.3	-4.7	4.7	–
6.9	-3.1	3.1	–
9.5	-0.5	0.5	–
11.5	1.5	1.5	+
10.1	0.1	0.1	+

Dari ilustrasi UKM sebelumnya:
Benarkah nilai tengah omset bulanan UKM Kota Bogor sebesar Rp 10 juta?

$H_0: M = 10$ vs $H_1: M \neq 10$

Ilustrasi

H_0: M = 10 vs H_1: M ≠ 10

			diurutkan			peringkat
8.1	-1.9	1.9		X	0	
7.8	-2.2	2.2		(+)	0.1	1
13.5	+3.5	3.5		(-)	0.5	2
12.1	+2.1	2.1		(+)	1.5	3
10.0	0	0		(-)	1.9	4
5.3	-4.7	4.7		(+)	2.1	5
6.9	-3.1	3.1		(-)	2.2	6
9.5	-0.5	0.5		(-)	3.1	7
11.5	+1.5	1.5		(+)	3.5	8
10.1	+0.1	0.1		(-)	4.7	9

W- = 2+4+6+7+9 = 28
W+ = 1+3+5+8 = 17
W = W+ = 17

Jumlah peringkat

$$z = \frac{\left| W - \frac{n(n+1)}{4} \right| - 0.5}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = 0.5923$$

- Nilai-p = 2*P(Z > 0.5923) = 0.554
- Bila taraf nyata α = 5%, maka H0 tidak ditolak → **belum cukup bukti** untuk menyatakan nilai tengah omset bulanan UKM Kota Bogor **tidak** sebesar Rp 10 juta

UJI JUMLAH PERINGKAT WILCOXON – MANN WHITNEY

Ide Uji Mann-Whitney

- Terkait uji nilai tengah dua populasi dengan penarikan contoh saling bebas
- Bila tidak ada perbedaan nilai tengah antara dua populasi,
  - maka data contoh kedua populasi tidak ada kecenderungan salah satunya lebih besar dari yang lain
  - Bila data contoh kedua populasi disatukan dan diberi peringkat, maka jumlah peringkat kedua gugus relatif sama besar

Ilustrasi

- Suatu kajian ingin membandingkan omset UKM di Kota Bogor dan Kabupaten Bogor. Ditengarai, tidak ada perbedaan nilai tengah omset UKM antara kedua wilayah ini
- Untuk keperluan ini, diambil 10 contoh UKM dari Kota Bogor dan 15 contoh UKM dari Kabupaten Bogor
- Besarnya omset masing-masing contoh UKM tersebut (juta rupiah) tersaji di tabel berikut ini

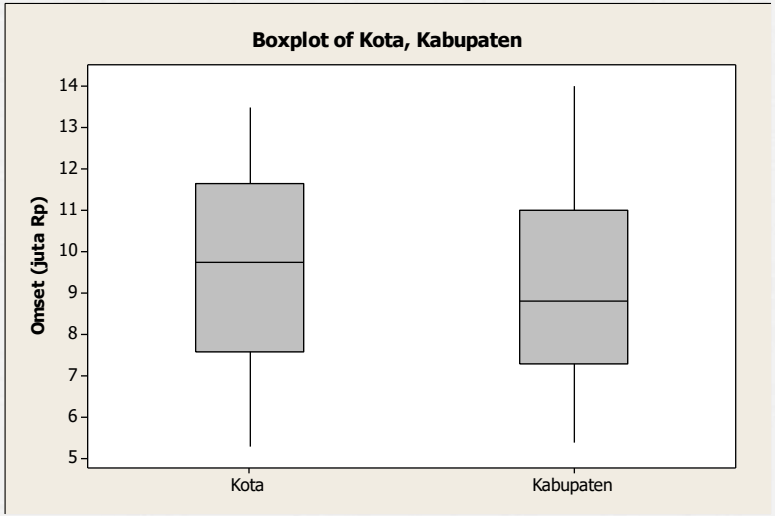
Kota Bogor	Kabupaten Bogor	
8.1	10.1	12.2
7.8	7.3	8.8
13.5	6.5	7.5
12.1	7.7	6.4
10.0	8.1	9.1
5.3	14.0	
6.9	13.0	
9.5	5.4	
11.5	9.3	
10.1	11.0	

M1 dan M2 masing-masing adalah median dari populasi 1 dan 2

→ Misalkan populasi 1 dan 2 masing-masing adalah UKM Kota Bogor dan Kabupaten Bogor

$H_0: M_1 = M_2$

$H_1: M_1 \neq M_2$



Prosedur pengujian

- Data dari kedua contoh digabung
- Diberikan peringkat
- Dijumlahkan peringkat dari amatan yang berasal dari populasi pertama → W

Prosedur pengujian

$H_0: M_1 \leq M_2$
$H_1: M_1 > M_2$

$$Z_w = \frac{W - \frac{n(m+n+1)}{2} - 0.5}{\sqrt{\frac{mn(m+n+1)}{12}}}$$

W = statistik uji Mann Whitney
n = # contoh dari populasi 1
m = # contoh dari populasi 2

$$p = P(Z > Z_w)$$

$H_0: M_1 \geq M_2$
$H_1: M_1 < M_2$

$$Z_w = \frac{S - \frac{n(m+n+1)}{2} - 0.5}{\sqrt{\frac{mn(m+n+1)}{12}}}$$

$$S = W - n(m+n+1)$$

$$p = P(Z < -Z_w)$$

Prosedur pengujian

H_0: $M_1 = M_2$
H_1: $M_1 \neq M_2$

$$Z_w = \frac{\left| W - \frac{n(m+n+1)}{2} \right| - 0.5}{\sqrt{\frac{mn(m+n+1)}{12}}}$$

W = statistik uji Mann Whitney
n = # contoh dari populasi 1
m = # contoh dari populasi 2

$$p = 2P(Z > Z_w)$$

Prosedur pengujian

Bila terdapat amatan yang bernilai sama, peringkat yang diberikan untuk mereka adalah rataan peringkat dan pembagi pada penghitungan Z_W di atas menjadi

$$\sqrt{\frac{mn}{12}\left[(m+n+1)-\frac{\sum_{i=1}^{I}\left(t_i^3-t_i\right)}{(m+n)(m+n-1}\right]}$$

i = 1, 2, ..., I

I = banyaknya set amatan yang bernilai sama

t_i = banyaknya amatan yang bernilai sama dari set amatan sama ke-1

No	Omset	Peringkat	Kelompok	No	Omset	Peringkat	Kelompok	No	Omset	Peringkat	Kelompok
1	5.3	1	Kota	12	8.1	10.5	Kabupaten	21	12.1	21	Kota
2	5.4	2	Kabupaten	12	8.8	12	Kabupaten	22	12.2	22	Kabupaten
3	6.4	3	Kabupaten	13	9.1	13	Kabupaten	23	13.0	23	Kabupaten
4	6.5	4	Kabupaten	14	9.3	14	Kabupaten	24	13.5	24	Kota
5	6.9	5	Kota	15	9.5	15	Kota	25	14.0	25	Kabupaten
6	7.3	6	Kabupaten	16	10.0	16	Kota				
7	7.5	7	Kabupaten	17	10.1	17	Kota				
8	7.7	8	Kabupaten	18	10.1	18	Kabupaten				
9	7.8	9	Kota	19	11.0	19	Kabupaten				
10	8.1	10.5	Kota	20	11.5	20	Kota				

W=138.5

- W = 138.5
- Zw = 0.4715
- p = 0.637
- Belum ada bukti cukup bahwa omset UKM dari kedua wilayah berbeda nyata

Uji Khi Kuadrat untuk Kebaikan Suai

Farit Mochamad Afendi
08128592194 – fmafendi@apps.ipb.ac.id

Pengujian Struktur Pasar

- Misalkan struktur pasar sebelumnya:
  Aqua = 40%, Vit = 30%, Nestle = 30%
- Aqua kemudian berinovasi dengan menciptakan varian minuman air putih dengan rasa buah.
- Beberapa waktu setelah peluncuran varian ini, dilakukan survei pasar yang melibatkan 200 responden. Masing-masing ditanyakan produk air minum dalam kemasan yang biasa digunakan.
- Hasilnya: pengguna Aqua 95, Vit 65, Nestle 40 orang.
- Apakah ada perubahan struktur pasar sebagai akibat dari inovasi ini?

Uji Khi kuadrat untuk kebaikan suai

- Kebutuhan di atas dapat dijawab menggunakan uji khi kuadrat (χ^2).
- Uji ini berbasis pada frekuensi amatan yang dibandingkan dengan frekuensi sesuai konteks yang ingin diuji

Uji Khi kuadrat untuk kebaikan suai

H0: tidak ada perubahan struktur pasar
H1: ada perubahan struktur pasar

$$\chi^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} \sim \chi^2_{(db=K-1)}$$

$$p = P\left(\chi^2 > \chi^2_{(db=K-1)}\right)$$

O_k = besarnya frekuensi teramati
E_k = besarnya frekuensi di bawah H0
K = banyaknya kategori sesuai konteks permasalahan

Uji Khi kuadrat untuk kebaikan suai

Merek	Share awal (%)	Frek survei	Frek bila tidak ada perubahan
Aqua	40	95	80
Vit	30	65	60
Nestle	30	40	60
		200	200
		O_k	E_k

$$\chi^2 = \frac{(95-80)^2}{80} + \frac{(65-60)^2}{60} + \frac{(40-60)^2}{60} = 9.896$$

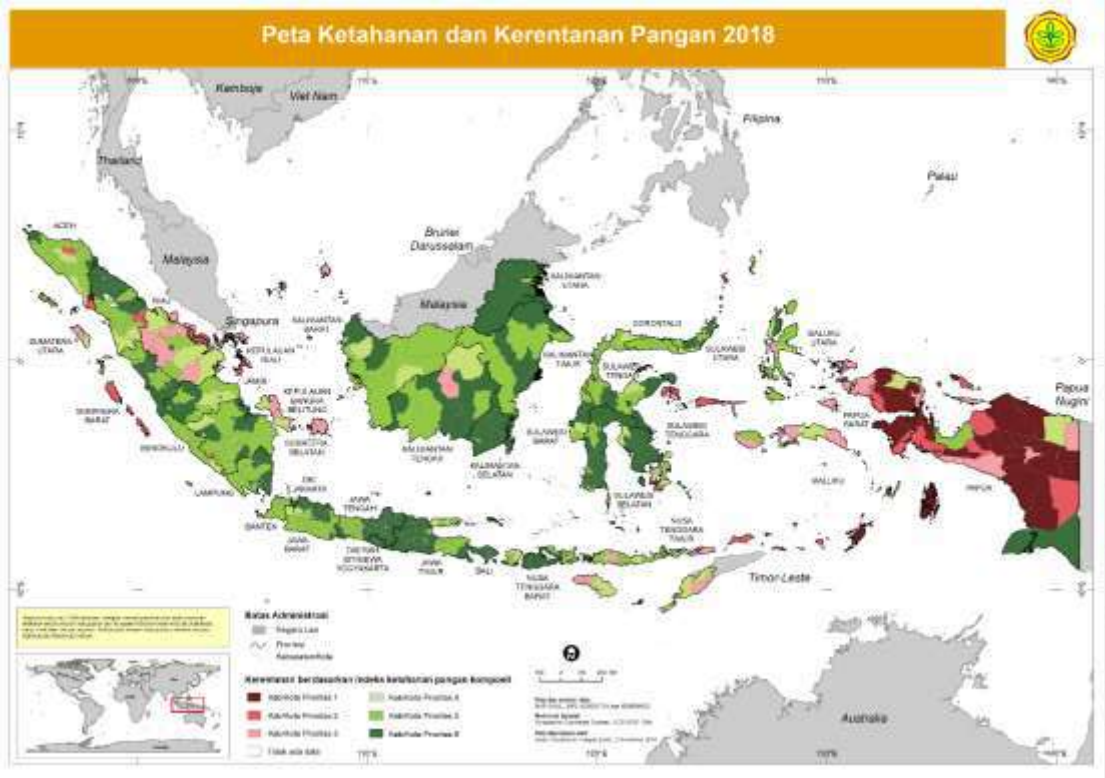
p = 0.007

Pada taraf nyata α=5%, tolak H0 → telah terjadi perubahan struktur pasar

STK351
Pengantar Analisis Data Kategorik

Farit Mochamad Afendi
08128592194 – fmafendi@apps.ipb.ac.id

Data kategorik di mana-mana

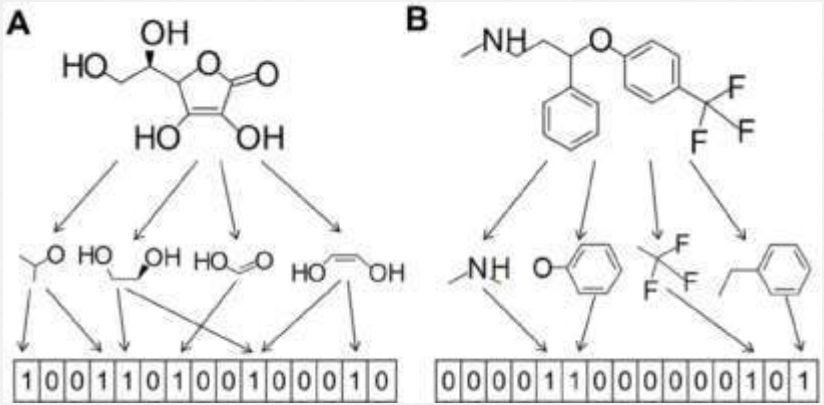


Data kategorik di mana-mana

Desain Obat: Kemiripan struktur geometris obat

Fingerprint	Abbreviation	Hashed	Length
EState fingerprint [24]	estate	NO	79
MACCS fingerprint [25]	maccs	NO	166
PubChem fingerprint [18]	pubchem	NO	881
Substructure fingerprint [18]	substructure	NO	308
Klekota Roth fingerprint [9]	KRFP	NO	4860
Fingerprint [26]	fingerprint	YES	1024
Extended fingerprint [18]	extended	YES	1024
Graph-only fingerprint [18]	graph only	YES	1024

doi:10.1371/journal.pone.0146666.t003



Struk belanja jadi uang?

https://www.kompasiana.com/salehafito/5bee9c25aeebe10e4e5d7f85/baru-tahu-saya-struk-bisa-jadi-uang

Transaction ID	Grapes	Apple	Mango	Orange
1	1	1	1	1
2	1	0	1	1
3	0	0	1	1
4	0	1	0	0
5	1	1	1	1
6	1	1	0	1

Market Basket Analysis
Recommendation Engine

Berita Terkait

- Hati-Hati dengan Bola Mati Inggris, Kroasia!
- 'Sepakbola Rusia Akan Mulai Dipercaya dan Dicintai'
- Top Skor Piala Dunia 2018: Kane Masih Memimpin
- Mimpi Brasil di Piala Dunia 2018 Terhenti, tapi Tidak Lenyap Sama Sekali
- Hasil Pertandingan Piala Dunia 2018: Rusia vs Kroasia Skor 2-2 (Adu Penalti 3-4)
- Pelatih Swedia Yakin Inggris Mampu Juara
- Hasil Pertandingan Piala Dunia 2018: Swedia vs Inggris Skor 0-2
- Henderson Masih Jadi Jimat Inggris

Baca Juga

Jordan Pickford Disebut Pendek, Berapa Tinggi Rata-rata Orang Inggris?

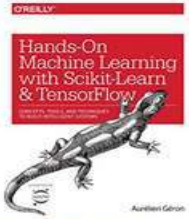
Dua Singa yang Antar Inggris ke Semifinal Piala Dunia Rusia

Video: Deretan Sepakan Indah di Perempat Final

Tarian Kemenangan Prancis atas Uruguay

Customers who bought this item also bought

Machine Learning with R - Second Edition: Expert techniques for predictive...
Brett Lantz
★★★★½ 85
Paperback
$48.69 prime



Hands-On Machine Learning with Scikit-Learn and TensorFlow:...
› Aurélien Géron
★★★★½ 208
#1 Best Seller in Artificial Intelligence
Paperback
$28.95 prime

Recommender Systems: The Textbook
› Charu C. Aggarwal
★★★★★ 7
Hardcover
$50.88 prime

Sponsored products related to thi

People also Bought

Taffware CLASSIC Velcro 15 - Black
Rp. 19.900,-

USB OTG Cable Multifunction Mobile Phone...
Rp. 6.300,-

Apple iPod Earphones (Original)
Rp. 32.900,-

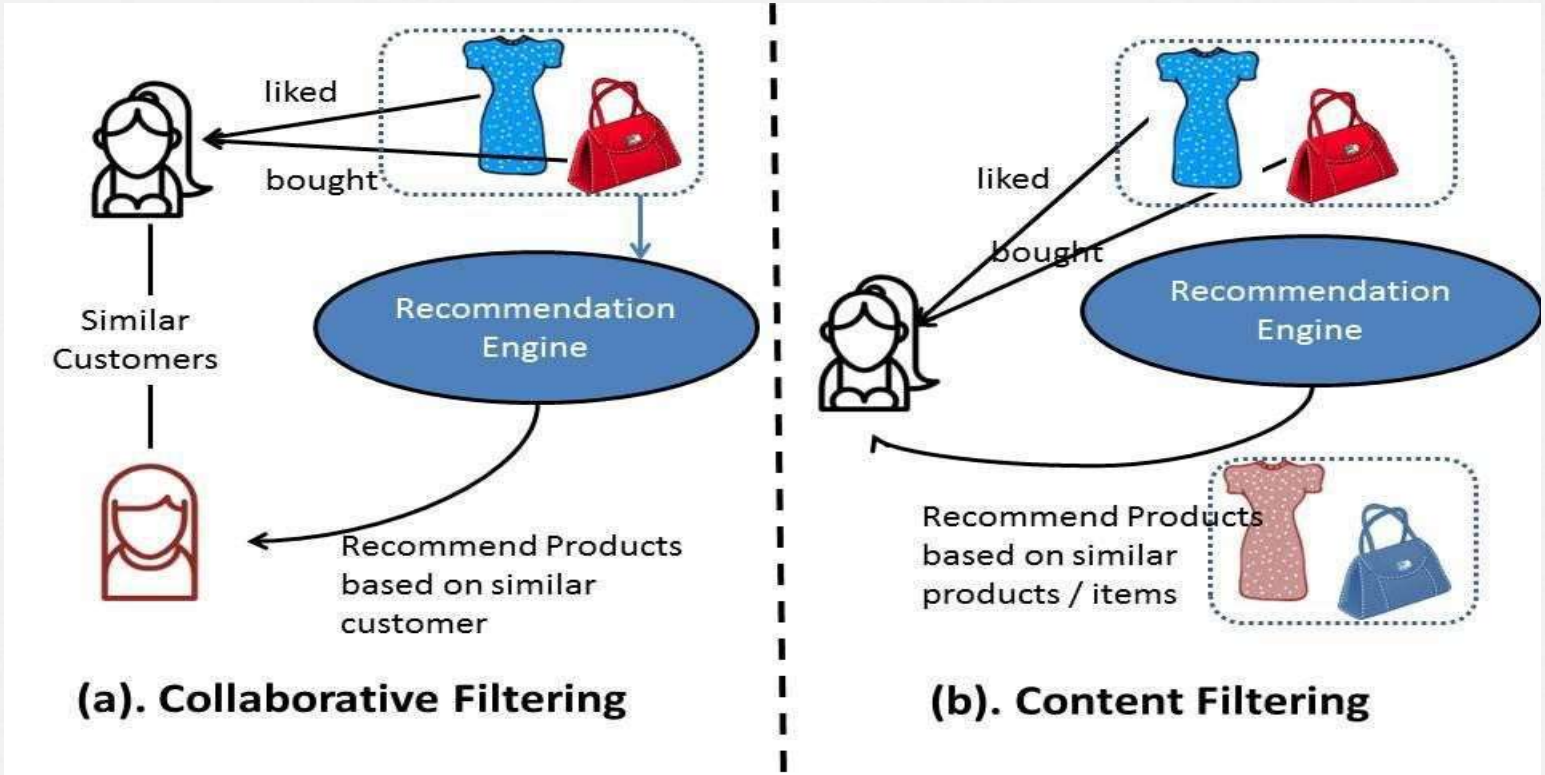
SanDisk Cruzer Blade USB Flash Drive 8GB...
Rp. 61.300,-

COD EXPRESS
Hari ini PESAN
Hari ini TERIMA

Related Items

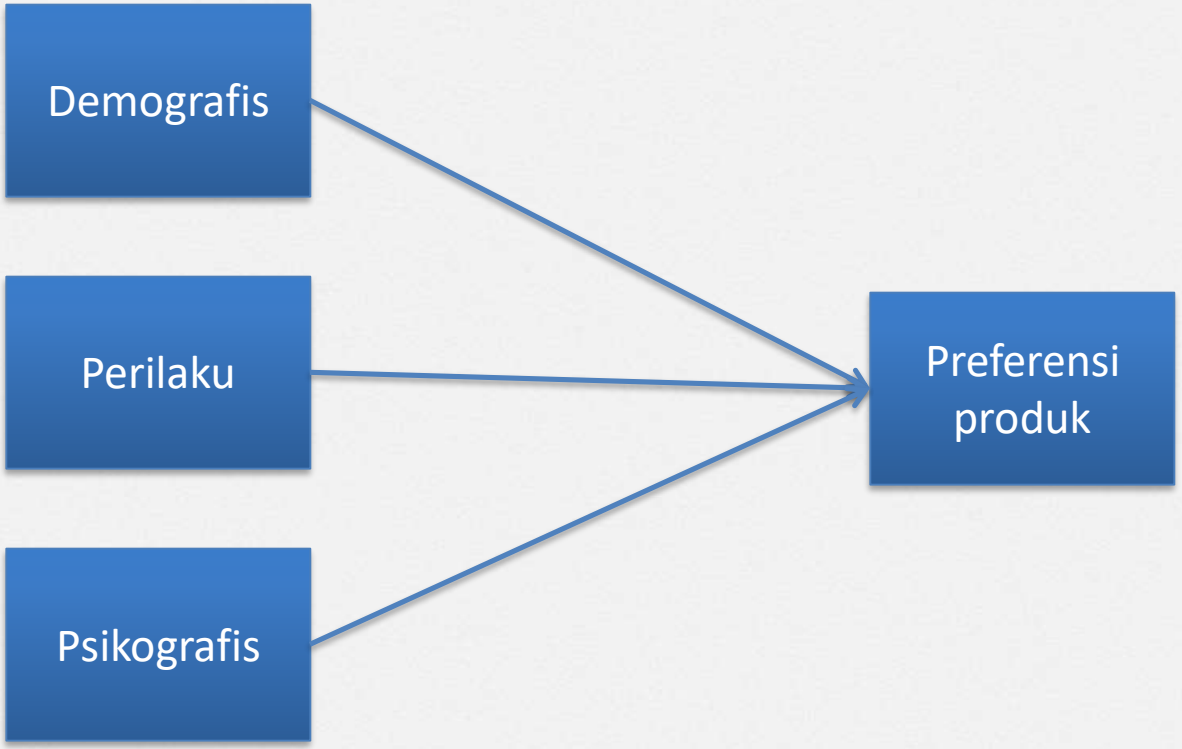
Soft Sleeve Case for Laptop 13 Inch - Gray
Rp. 24.700,-

Berbagai tipe RE



Preferensi produk

Peran peubah



BASIS SEBARAN

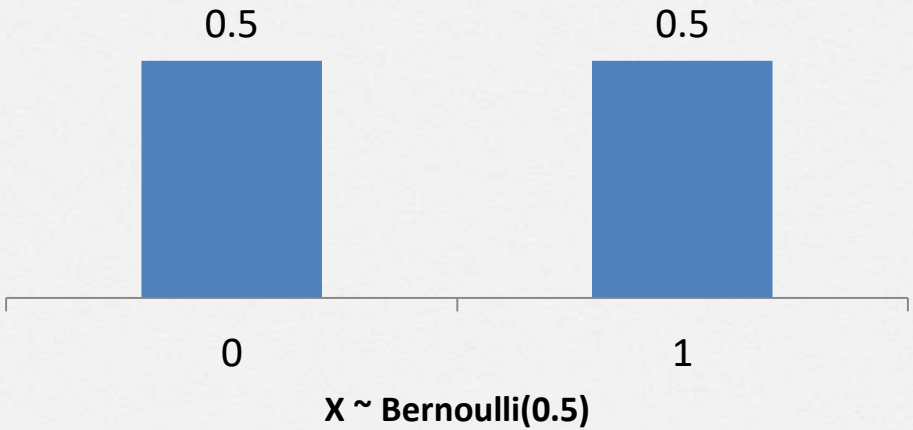
Peubah acak Bernoulli

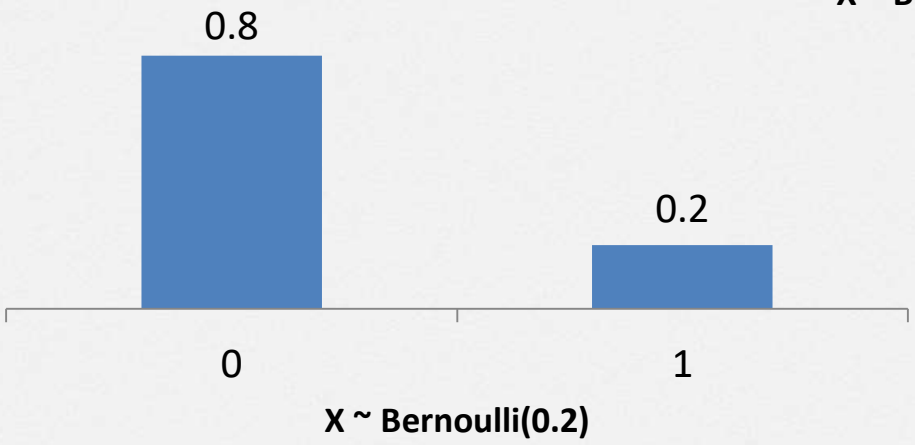
- Berkenaan dengan fenomena dengan dua kemungkinan hasil:
  - sukses → X = 1
  - Gagal → X = 0
- Peluang sukses = p
- Fungsi peluang

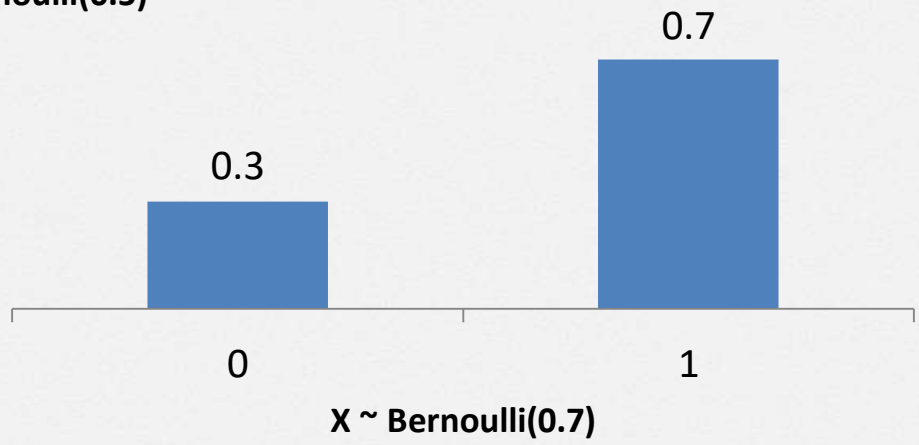
$$P(X = x) = \begin{cases} p^x (1-p)^{1-x} \text{ ; untuk } x = 0, 1 \\ 0 \text{ ; untuk } x \text{ lainnya} \end{cases}$$

- X ~ Bernoulli(p)

Ilustrasi Sebaran Bernoulli





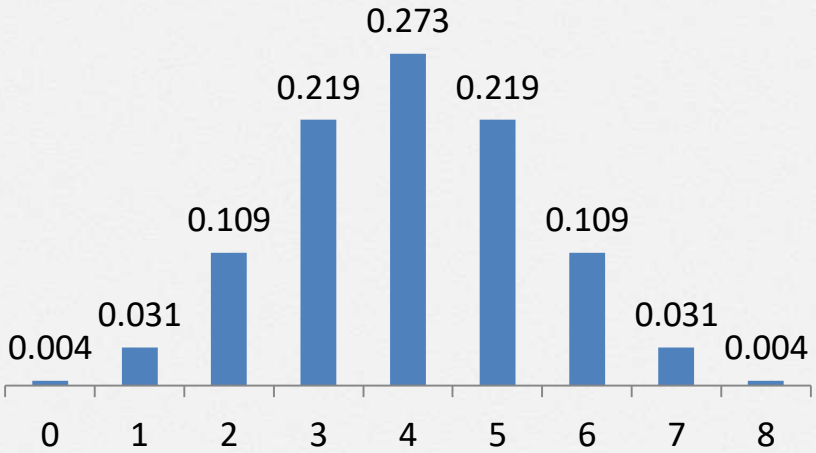


Peubah acak Binom

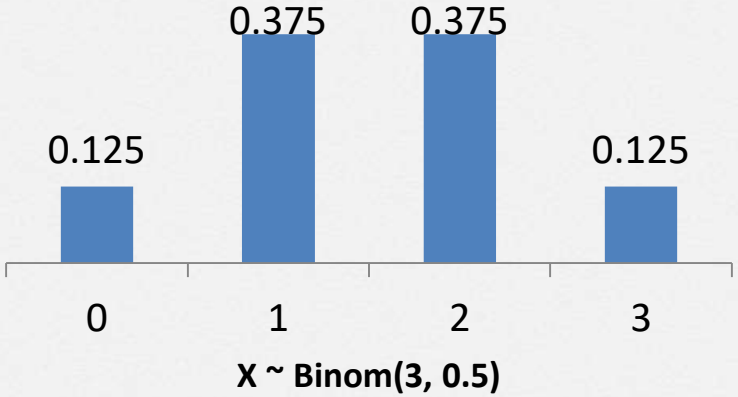
- Jumlah kejadian sukses dari n kejadian Bernoulli yang saling bebas dengan peluang sukses tetap sebesar p
- X ~ Binom(n, p)
- Fungsi peluang

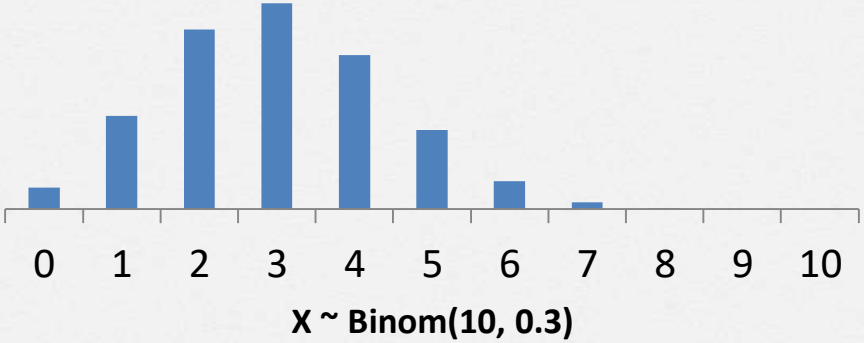
$$P(X=x)=\begin{cases}\binom{n}{x}p^x(1-p)^{n-x}\ ;\ \text{untuk } x=0,1,2,\ldots,n\\ 0\ ;\ \text{untuk } x \text{ lainnya}\end{cases}$$

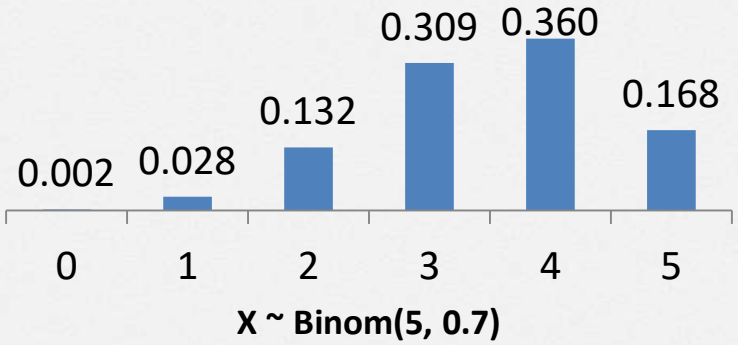
- Pengaturan 0 dan 1 menjadi bilangan dengan 8 dijit
- Banyaknya angka 1 pada bilangan yang terbentuk → X ~ Binom(8, 0.5)

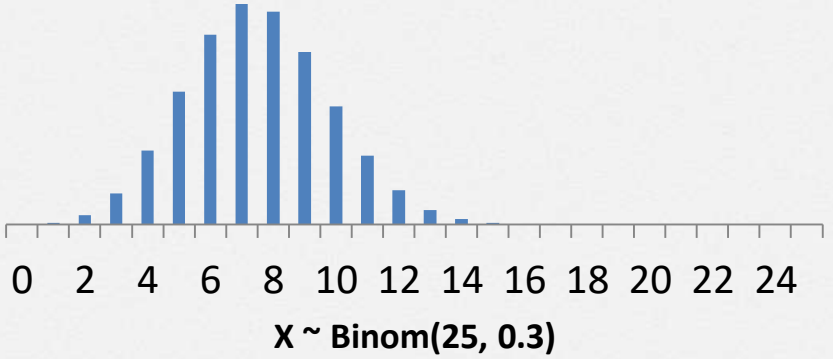


Ilustrasi Sebaran Binom







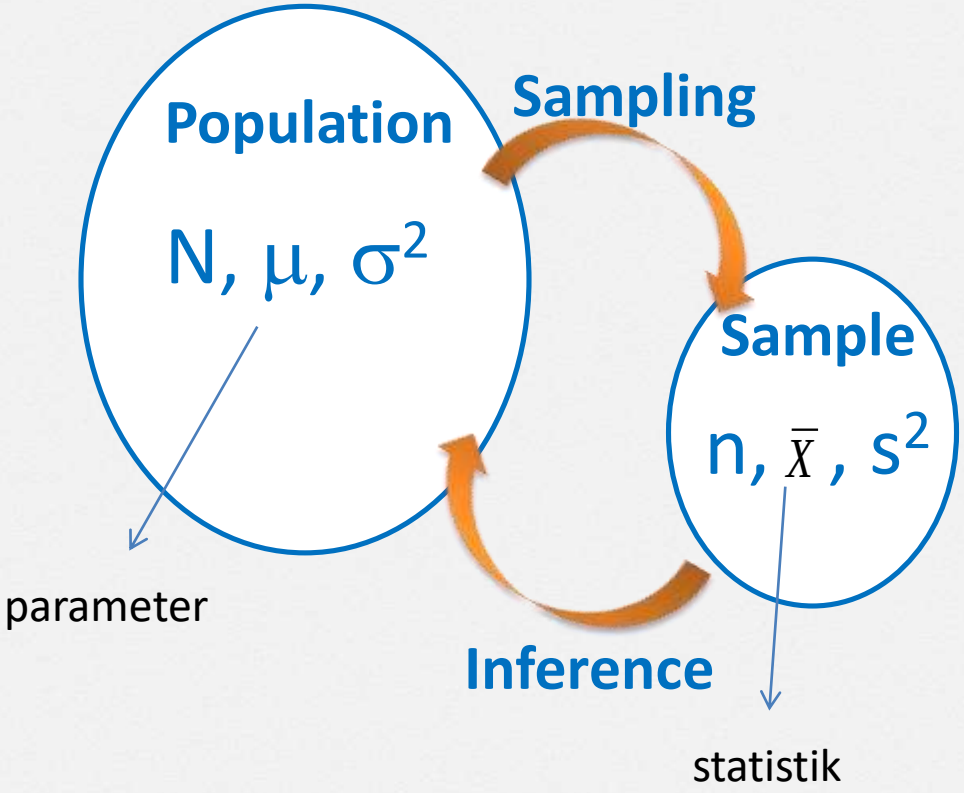


Sebaran Multinomial

- Serupa dengan Sebaran binomial, hanya “pilihan kejadiannya” lebih dari dua.

$$P(n_1, n_2, \ldots, n_c) = \left(\frac{n!}{n_1!n_2!\cdots n_c!}\right)\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_c^{n_c}$$

Dari Contoh ke Populasi



Inferensia Statistika:

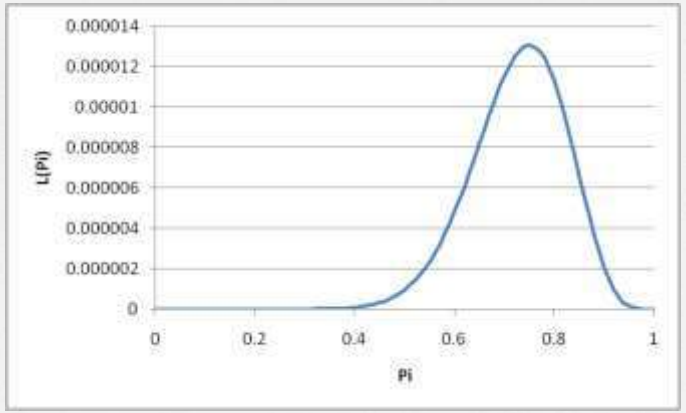
- Pendugaan
  - Tidak ada asumsi/hipotesis tentang populasi
  - Peran contoh menyediakan informasi mengenai populasi
- Pengujian hipotesis
  - Ada asumsi/hipotesis yang disusun berkaitan dengan populasi
  - Peran contoh menyediakan bukti keberlakuan asumsi/hipotesis tersebut

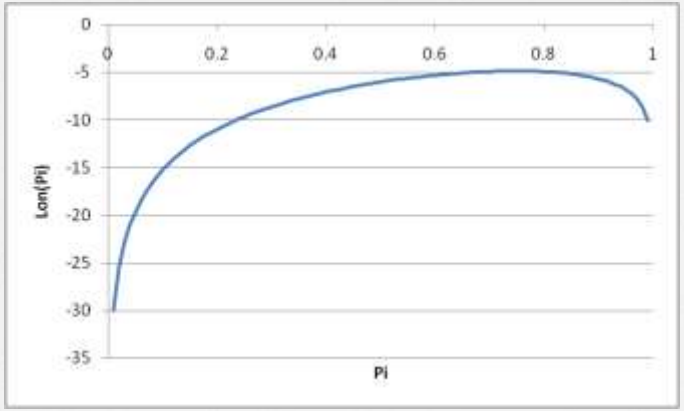
Pendugaan parameter

- Performa suatu merek di benak konsumen (*brand awareness*) sering digali dalam riset pemasaran
- Bentuknya:
  - *Top of mind*: ingat langsung tanpa dibantu
  - *Aided*: ingat dengan dibantu

Pendugaan parameter

- Survei pada 20 orang konsumen kecap menghasilkan 15 orang di antaranya ingat merek “A” pada penyebutan pertama tanpa bantuan.
- Berapa ToM kecap merek A tersebut di level populasi?





Pengujian Proporsi

- Produsen Aqua mengklaim memimpin pasar dengan share 40%.
- Misalkan dari survei terhadap 90 konsumen, 30 di antaranya membeli Aqua.
- Apakah klaim produsen Aqua dapat diterima?

Langkah-langkah pengujian

1. Klaim share 40% → $\pi = 0.4$
2. H0: $\pi = 0.4$ vs H1: $\pi \neq 0.4$
3. Survei: n = 90, X = 30 → p = 30/90 = 0.333
4. n besar → Uji Z;

$$Z = \frac{(p - \pi_0)}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Test and CI for One Proportion

Test of p = 0.4 vs p not = 0.4

Sample	X	N	Sample p	95% CI	Z-Value	P-Value
1	30	90	0.333333	(0.235942, 0.430725)	-1.29	0.197

Using the normal approximation.

Pengujian Proporsi (contoh kecil)

- Produsen Aqua mengklaim memimpin pasar dengan share 40%.
- Misalkan dari survei terhadap 10 konsumen, 3 di antaranya membeli Aqua.
- Apakah klaim produsen Aqua dapat diterima?

Langkah-langkah pengujian

1. Klaim share 40% → $\pi = 0.4$
2. H0: $\pi = 0.4$ vs H1: $\pi \neq 0.4$
3. Survei: n = 10, X = 3 → p = 3/10 = 0.3
4. n kecil → Uji berbasis binomial

```
Test and CI for One Proportion

Test of p = 0.4 vs p not = 0.4

                                                                Exact
Sample  X   N  Sample p          95% CI                      P-Value
1       3  10  0.300000  (0.066740, 0.652453)                 0.549
```

Pengujian Struktur Pasar

- Misalkan struktur pasar sebelumnya:
  Aqua = 40%, Vit = 30%, Nestle = 30%
- Aqua kemudian berinovasi dengan menciptakan varian minuman air putih dengan rasa buah.
- Beberapa waktu setelah peluncuran varian ini, dilakukan survei pasar yang melibatkan 200 responden. Masing-masing ditanyakan produk air minum dalam kemasan yang biasa digunakan.
- Hasilnya: pengguna Aqua 95, Vit 65, Nestle 40 orang.
- Apakah ada perubahan struktur pasar sebagai akibat dari inovasi ini?

Uji Khi kuadrat untuk kebaikan suai

- Kebutuhan di atas dapat dijawab menggunakan uji khi kuadrat (χ^2).
- Uji ini berbasis pada frekuensi amatan yang dibandingkan dengan frekuensi sesuai konteks yang ingin diuji

Uji Khi kuadrat untuk kebaikan suai

H0: tidak ada perubahan struktur pasar
H1: ada perubahan struktur pasar

$$\chi^2 = \sum_{k=1}^{K} \frac{\left(O_k - E_k\right)^2}{E_k} \sim \chi^2_{(db=K-1)}$$

$$p = P\left(\chi^2 > \chi^2_{(db=K-1)}\right)$$

O_k = besarnya frekuensi teramati
E_k = besarnya frekuensi di bawah H0
K = banyaknya kategori sesuai konteks permasalahan

Uji Khi kuadrat untuk kebaikan suai

Merek	Share awal (%)	Frek survei	Frek bila tidak ada perubahan
Aqua	40	95	80
Vit	30	65	60
Nestle	30	40	60
		200	200
		O_k	E_k

$$\chi^2 = \frac{(95-80)^2}{80} + \frac{(65-60)^2}{60} + \frac{(40-60)^2}{60} = 9.896$$

p = 0.007

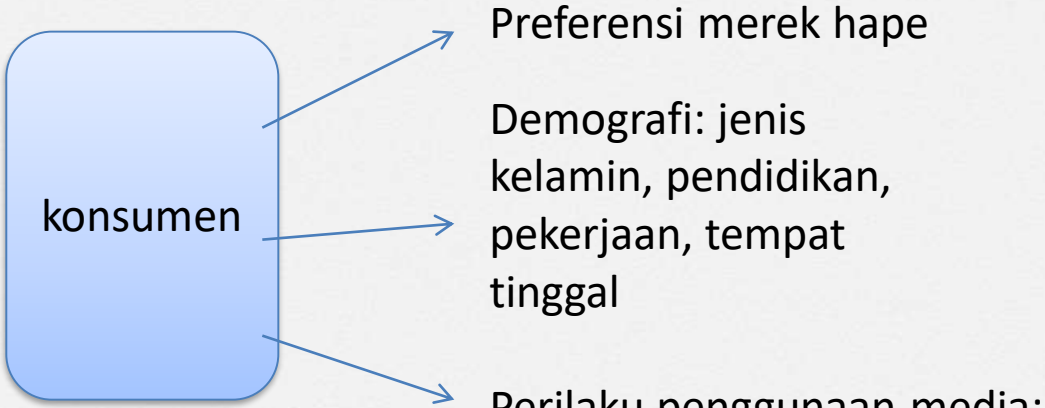
Pada taraf nyata α=5%, tolak H0 → telah terjadi perubahan struktur pasar

STK351
Pengantar Analisis Data Kategorik

Farit Mochamad Afendi
08128592194 – fmafendi@apps.ipb.ac.id

Tabulasi

Suatu survei dilakukan untuk mendapatkan gambaran preferensi terhadap merek hape tertentu. Dari survei ini diharapkan diperoleh gambaran profil konsumen yang cenderung memilih merek hape tertentu. Informasi ini akan sangat bermanfaat untuk penajaman *marketing campaign*, desain produk, dsb.



Tabulasi

merek Hape	Frek
A	85
B	115
C	200

Jenis Kelamin	Frek
Pria	220
Wanita	180

Jenis Kelamin	merek Hape			Total
	A	B	C	
Pria	35	25	160	220
Wanita	50	90	40	180
Total	85	115	200	400

Tabel kontingensi (Pearson)
Tabulasi silang

Struktur peluang

Jenis Kelamin	merek Hape			Total
	A	B	C	
Pria	35	25	160	220
Wanita	50	90	40	180
Total	85	115	200	400

n_{ij} n_{i+} n_{+j} n_{++}

Sebaran bersama $\pi_{ij} = n_{ij} / n$

Sebaran marjinal $\pi_{i+} = n_{i+} / n$

$\pi_{+j} = n_{+j} / n$

n acak → Y_{ij} ~ Poisson(μ_{ij})

n tetap → Y_{ij} ~ Multinomial(n,π_{ij}); $\pi_{ij} = \mu_{ij}/n$

n_i tetap → Y_{ij} ~ Multinomial(n_i,$\pi_{j|i}$); $\pi_{ij} = \mu_{ij}/\mu_i$

n_i dan n_j tetap → hipergeometrik

Struktur peluang

Row	Column 1	Column 2	Total
1	π_{11} $(\pi_{1\|1})$	π_{12} $(\pi_{2\|1})$	π_{1+} (1.0)
2	π_{21} $(\pi_{1\|2})$	π_{22} $(\pi_{2\|2})$	π_{2+} (1.0)
Total	π_{+1}	π_{+2}	1.0

Saling bebas: $\pi_{j|i} = \pi_{ij}/\pi_{i+} = (\pi_{i+}\,\pi_{+j})/\pi_{i+} = \pi_{+j}$

Tipe studi

- *Retrospective*: kasus diamati di masa kini, ditelusuri peristiwa yang terjadi di masa lalu
  - *Case control*
- *Prospective*: kondisi diamati sekarang untuk diamati dampaknya di masa depan
  - *Clinical trial* (alokasi perlakuan acak)
    - *cohort study* (alokasi perlakuan sukarela)
  - *Cross sectional study* (contoh dipilih untuk diamati perlakuan dan respon sekaligus)

Tipe studi

- *Restrospective study*
  - mengendalikan n+j,
  - menganggap frekuensi I sebagai contoh dari sebaran multinomial
- *Prospective study*
  - mengendalikan ni+,
  - menganggap frekuensi J sebagai contoh dari sebaran multinomial
- *Cross sectional study*
  - mengendalikan n,
  - menganggap frekuensi IJ sebagai contoh dari sebaran multinomial

Tipe studi

- Observational study
  - Case control
  - Cohort
  - Cross sectional
- Experimental study
  - Clinical trial

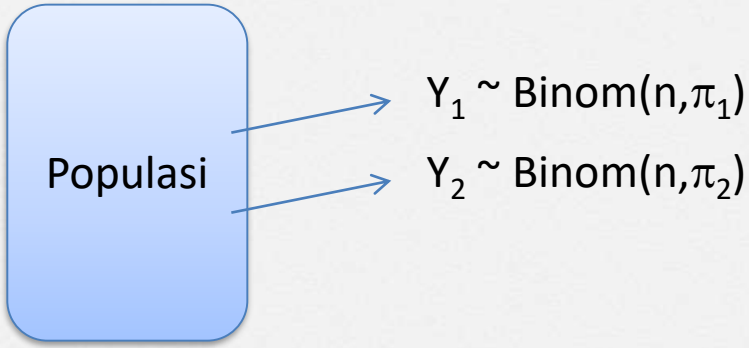
Kanker	Hasil diagnosa		Total
	Positif	Negatif	
Ya	85	15	100
Tidak	50	150	200
Total	135	165	300

Kepekaan (*sensitivity*) → kemampuan mendeteksi yang sakit

85/100 = 0.85

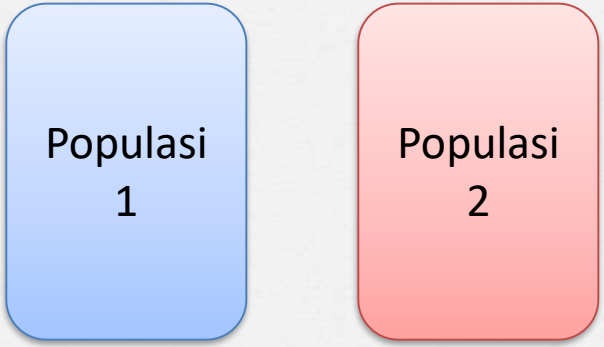
Kekhususan (*specificity*) → kemampuan mendeteksi yang tidak sakit

150/200 = 0.75



Pengujian asosiasi: $\pi_{ij} = \pi_i \pi_j$

$E_{ij} = \pi_i \pi_j n$
$= (n_i/n)(n_j/n) n$
$= (n_i n_j/n)$



$Y \sim Binom(n,\pi_1)$ $Y \sim Binom(n,\pi_2)$

Pengujian kehomogenan: $\pi_1 = \pi_2 = \pi$

$E_{ij} = n_i \pi$
$= n_i n_j/n$

Uji Khi Kuadrat

$$\chi^2 = \sum_{i=1}^{b}\sum_{j=1}^{k}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(db=bk-1)}$$

$$p = P\left(\chi^2 > \chi^2_{(db=bk-1)}\right)$$

O_{ij} = besarnya frekuensi teramati
E_{ij} = besarnya frekuensi di bawah H_0
b = banyaknya baris
k = banyaknya kolom

Jenis Kelamin	merek Hape			Total
	A	B	C	
Pria	35 (46.75)	25 (63.25)	160 (110)	220
Wanita	50 (38.25)	90 (51.75)	40 (90)	180
Total	85	115	200	400

H0: Tidak ada asosiasi antara jenis kelamin dan preferensi merek HP
H1: Ada asosiasi antara jenis kelamin dan preferensi merek HP

E11 = (220 . 85)/400 = 46.75
E21 = (180 . 85)/400 = 38.25

$$\chi^2 = \frac{(35-46.75)^2}{46.75} + \ldots + \frac{(40-90)^2}{90} = 108.471$$

$$p = P(\chi^2_{db=4} > 108.471) = 0.000$$

Tolak H0 → ada asosiasi antara keduanya

Beda Proporsi

During the early 1950s, polio rates in the U.S. were above 25,000 annually; in 1952 and 1953, the U.S. experienced an outbreak of 58,000 and 35,000 polio cases, respectively, up from a typical number of some 20,000 a year, with deaths in those years numbering 3,200 and 1,400.

"Polio pioneers"—some of the many children who took part in trials of poliomyelitis vaccine

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1114166/

The first effective polio vaccine was developed in 1952 by Jonas Salk and a team at the University of Pittsburgh that included Julius Youngner, Byron Bennett, L. James Lewis, and Lorraine Friedman, which required years of subsequent testing.

Beda Proporsi

$$H_0 : \pi_1 = \pi_2$$
$$H_0 : \pi_1 \neq \pi_2$$

- Populasi 1: mendapat vaksin
- Populasi 2: mendapat plasebo
- n_1 dan n_2: banyaknya contoh dari populasi 1 dan 2
- x_1 dan x_2: banyaknya kasus dari contoh populasi 1 dan 2
- π_1 dan π_2: proporsi populasi 1 dan 2
- p_1 dan p_2: proporsi contoh populasi 1 dan 2

$$z = \frac{(p_1 - p_2)}{SE}$$

$$p_1 = \frac{x_1}{n_1} \qquad p_2 = \frac{x_2}{n_2}$$

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Beda Proporsi

Double blind experiment

Study group	Study population	Total Polio Case
Vaccinated	200745	57
Placebo	201229	142


```
Test and CI for Two Proportions

Sample    X       N  Sample p
1        57  200745  0.000284
2       142  201229  0.000706

Difference = p (1) - p (2)
Estimate for difference:  -0.000421721
95% CI for difference:  (-0.000559175, -0.000284267)
Test for difference = 0 (vs not = 0):  Z = -6.01  P-Value = 0.000

Fisher's exact test: P-Value = 0.000
```

Perbandingan proporsi

Jenis kelamin	merek hape		Total
	A	B	
Pria	35	185	220
Wanita	50	130	180
Total	85	315	400

Resiko relatif

P(A|Pria) = 35/220 = 0.16

P(A|Wanita) = 50/180 = 0.28

RR = 0.16/0.28 = 0.57

Rasio Odds

P(A|Pria) = 35/220 = 0.16

P(B|Pria) = 185/220 = 0.84

Odds pria = 0.16/0.84 = 0.19

P(A|Wanita) = 50/180 = 0.28

P(B|Wanita) = 130/180 = 0.72

Odds wanita = 0.28/0.72 = 0.38

Rasio odds = 0.19/0.38 = 0.49