METODE STATISTIKA

Penulis:
A'yunin Sofro
Affiati Oktaviarina
Dimas Avian Maulana

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI SURABAYA

KATA PENGANTAR

Alhamdulillahirrabilalamin, segala puji dan syukur kehadirat Allah SWT, yang telah melimpahkan rahmat-Nya didalam proses penulisan buku ini. Penulis ingin mengucapkan terima kasih kepada mereka yang telah memberikan bantuan dan dukungan mereka selama proses menulis; menuju penyelesaian buku ini pada khususnya.

Terima kasih kepada reviewer, Dr. Ismail, M.Pd atas bimbingan dan kesabarannya selama penulisan ini. Kesempatan untuk belajar darinya sangat berharga dan tidak akan dilupakan. Penulis berhutang budi kepada Universitas Negeri Surabaya, Indonesia karena kesempatannya yang diberikan untuk menulis buku ini.

Terima kasih khusus saya kepada suami penulis Sony Febrianto, anak-anak penulis Zaky Athallah Muhammad dan Natasya Azzahra atas kesabaran dan dukungan mereka. Tidak lupa, orang tua dan anggota keluarga lainnya atas dukungan moral mereka selama ini. Terakhir dan tidak kalah pentingnya, untuk semua teman terutama tim yang telah membantu penulis menjalani masa-masa menulis bersama, yang semuanya bermanfaat untuk keberhasilan terbitnya buku ini, – A. S

Terima kasih banyak kepada keluarga atas kesabaran dan dukungan baik moril dan material yang diberikan kepada penulis. Terimakasih juga saya sampaikan kepada tim yang telah bersusah payah memberikan pelatihan Latex kepada penulis, — A. O

Terima kasih banyak seluruh pihak yang tidak bisa penulis sebutkan satu persatu. Terimakasih juga saya sampaikan kepada tim yang saling support dalam penulisan buku ajar ini. Semoga buku ini dapat berguna bagi para pembaca sekalian — D. A. M.

Surabaya, 26 Agustus 2019

Penulis

DAFTAR ISI

HALAMAN JUDUL i

KATA PENGANTAR iii

DAFTAR ISI v

DAFTAR GAMBAR ix

DAFTAR TABEL xi

DAFTAR *SOURCE CODE* xiii

BAB I Pengetahuan Dasar Statistika 1
- A. Pendahuluan 1
  - 1. Deskripsi Singkat 1
  - 2. Tujuan Pembelajaran 1
- B. Materi 1
  - 1. Statistik dan Statistika 1
  - 2. Populasi dan Sampel 2
  - 3. Teknik Sampling 3
  - 4. Variabel dan Data 3
  - 5. Rangkuman 5
- C. Perlatihan 5
- D. Bahan Bacaan 6

BAB II Ukuran Pemusatan dan Penyebaran Data 7
- A. Pendahuluan 7
  - 1. Deskripsi Singkat 7
  - 2. Tujuan Pembelajaran 7
- B. Materi 7
  - 1. Ukuran Pemusatan Data 7
  - 2. Ukuran Penyebaran Data 9
  - 3. Rangkuman 13
- C. Perlatihan 14
- D. Buku Bacaan 14

BAB III Visualisasi Data 15
A. Pendahuluan 15
1. Deskripsi Singkat 15
2. Tujuan Pembelajaran 15
B. Materi 15
1. Histogram 15
2. Box-plot 18
3. Rangkuman 20
C. Perlatihan 20
D. Buku Bacaan 21

BAB IV Ruang Sampel, Peluang dan Distribusi Peluang 23
A. Pendahuluan 23
1. Deskripsi Singkat 23
2. Tujuan Pembelajaran 23
B. Materi 23
1. Ruang Sampel 23
2. Peluang dari suatu kejadian 26
3. Variabel Acak dan Distribusi Peluang 28
4. Distribusi Normal 30
5. Rangkuman 33
C. Perlatihan 33
D. Bahan Bacaan 34

BAB V Uji Hipotesis, Uji Z dan Uji T 35
A. Pendahuluan 35
1. Deskripsi Singkat 35
2. Tujuan Pembelajaran 35
B. Materi 35
1. Uji Hipotesis 35
2. Uji Z dan uji T pada satu sampel 40
3. Uji Z dan uji T pada dua sampel 46
4. Rangkuman 51
C. Perlatihan 52
D. Buku Bacaan 53

BAB VI Anova 55
A. Pendahuluan 55
1. Deskripsi Singkat 55
2. Tujuan Pembelajaran 55
3. Anova satu arah 55

4. Rangkuman 59
B. Perlatihan 59
C. Buku Bacaan 60

BAB VII Korelasi 61
A. Pendahuluan 61
1. Deskripsi Singkat 61
2. Tujuan Pembelajaran 61
B. Materi 61
1. Persamaan Korelasi 61
2. Diagram Pencar dan Korelasi 63
3. Koefisien Korelasi 65
4. Pengujian Hipotesis Koefisien Korelasi 68
5. Rangkuman 71
C. Perlatihan 71
D. Bahan Bacaan 72

BAB VIII Regresi 75
A. Pendahuluan 75
1. Deskripsi Singkat 75
2. Tujuan Pembelajaran 75
B. Materi 75
1. Persamaan Regresi 75
2. Regresi Linear Sederhana 77
3. Analisis residual 82
4. Pengujian Hipotesis Koefisien Regresi 85
5. Pengujian Hipotesis Konstanta Regresi 89
6. Rangkuman 91
C. Perlatihan 92
D. Bahan Bacaan 93

BAB IX Uji *Chi square* 95
A. Pendahuluan 95
1. Deskripsi Singkat 95
2. Tujuan Pembelajaran 95
B. Materi 95
1. Distribusi *Chi-Square* 95
2. Uji Kebaikan Suai 96
3. Uji Independensi *Chi-Square* 102
4. Rangkuman 104
C. Perlatihan 105

D. Bahan Bacaan 105
DAFTAR PUSTAKA 107
GLOSARIUM 107

DAFTAR GAMBAR

Gambar 3.1 Histogram untuk contoh 3.10 17
Gambar 3.2 Histogram untuk contoh 3.11 19
Gambar 3.3 Boxplot untuk contoh 3.12 19

Gambar 4.1 Kurva distribusi normal. 31
Gambar 4.2 Daerah dibawah kurva 31
Gambar 4.3 Kurva dari Distribusi Student's t. 32

Gambar 5.1 Distribusi dari rata rata sampel 36
Gambar 5.2 Daerah diterimanya dugaan awal (H_0)dari dua ekor (*two tails*) 38
Gambar 5.3 Daaerah kritis *two tails* dengan *significant level*= 0.05 38
Gambar 5.4 Daaerah kritis *two tails* dengan *significant level*= 0.01 39

Gambar 7.1 Diagram pencar lama belajar vs. nilai ujian . 65
Gambar 7.2 Rentang nilai koefisien korelasi 66
Gambar 7.3 Hubungan antara koefisien korelasi dan diagram pencar 66

Gambar 8.1 Suatu hubungan linear 76
Gambar 8.2 Diagram pencar dan garis regresi 80
Gambar 8.3 Residual dari suatu titik 84
Gambar 8.4 Kurva residual yang menunjukkan pelanggaran asumsi 84
Gambar 8.5 Kurva residual 86

Gambar 9.1 Kurva χ^2 untuk beberapa derajat kebebasan 96

DAFTAR TABEL

Tabel 1.1 Contoh Data Kualitatif dan Kuantitatif 5
Tabel 1.2 Contoh Data Nominal dan Ordinal 5

Tabel 2.1 Penyelesaian untuk contoh 2.5 9
Tabel 2.2 Penyelesaian untuk contoh 2.6 10
Tabel 2.3 Penyelesaian contoh 2.7 10
Tabel 2.4 Penyelesaian contoh 2.8 12

Tabel 3.1 Langkah-Langkah Membuat tabel frekuensi ... 16
Tabel 3.2 Langkah-Langkah Membuat Histogram 16
Tabel 3.3 Penyelesaian untuk contoh 3.10 16
Tabel 3.4 Tabel frekuensi untuk contoh 3.10 17
Tabel 3.5 Tabel suhu di 50 daerah 18

Tabel 4.1 Peluang dari X, banyaknya komponen yang cacat dari tiga komponen elektronik yang ditest 30

Tabel 5.1 Penyelesaian Uji T satu sampel dengan menggunakan pendekatan *P'value* 45
Tabel 5.2 Prosedur uji hipotesis pada statistika inferensial 52

Tabel 6.1 Kenaikan berat badan (oz) tikus memberi makan empat merek sereal yang berbeda 56
Tabel 6.2 Kenaikan berat badan (oz) tikus memberi makan empat merek sereal yang berbeda 56
Tabel 6.3 Anova satu arah 56
Tabel 6.4 Kenaikan berat badan (oz) tikus memberi makan empat merek sereal yang berbeda 57
Tabel 6.5 Anova satu arah 57
Tabel 6.6 Umur dari tiga komunitas yang berbeda 59
Tabel 6.7 kumulatif radiokatif dari tiga lokasi yang berbeda 60

Tabel 7.1 Lama belajar dan Nilai ujian 63

Tabel 8.1 Data penjualan mobil di Amerika Serikat pada tahun 2018 79
Tabel 8.2 Tabel lengkap data penjualan mobil 79
Tabel 8.3 Tabel residual dari data penjualan mobil 85

Tabel 9.1 Jenis kriminalitas 97
Tabel 9.2 Jenis kriminalitas dan frekuensinya 97
Tabel 9.3 Tabel kontingensi 102
Tabel 9.4 Hasil survei dari universitas X 104

DAFTAR *SOURCE CODE*

7.1 *Syntax* R untuk membuat diagram pencar 64
7.2 *Syntax* R untuk menghitung koefisien korelasi 68
8.1 *Syntax* R untuk membuat garis regresi 80
8.2 *Syntax* R untuk membuat kurva residual 85
9.1 *Syntax* R untuk uji *chi-square* frekuensi harapan sama 100
9.2 *Syntax* R untuk uji *chi-square* frekuensi harapan tidak sama 102
9.3 *Syntax* R untuk uji independensi *chi-square* 103

BAB I
PENGETAHUAN DASAR STATISTIKA

A. Pendahuluan

1. Deskripsi Singkat

Dalam bab ini akan dibahas pengertian tentang statistik, statistika, populasi dan sampel.

2. Tujuan Pembelajaran

a. Menjelaskan pengetahuan tentang statistik dan statistika
b. Menjelaskan pengetahuan tentang populasi dan sampel
c. Menjelaskan pengetahuan tentang teknik sampling
d. Menerapkan pengetahuan tentang teknik sampling dalam kehidupan sehari-hari
e. Menjelaskan pengetahuan tentang variabel dan data

B. Materi

1. Statistik dan Statistika

Statistik berasal dari bahasa Latin ″status″ yang berarti ″negara″. Istilah statistik telah digunakan sejak jaman dahulu untuk memberikan informasi di bidang ekonomi, demografis dan politik di suatu negara. Seiring bertambahnya waktu, setiap negara seakan berlomba-lomba dalam mengembangkan statistik. Setiap negara mendirikan lembaga khusus yang bertugas menangani statistik di semua sektor pemerintahan dan membagikan informasi yang diperoleh kepada masyarakat secara berkala. Informasi ini sangat bermanfaat bagi masyarakat untuk memberikan transparansi sistem pemerintahan yang sedang dijalankan dan juga untuk mendukung peneliti untuk mengembangkan penelitian di bidang tersebut. Di Indonesia, terdapat Badan Pusat Statistik (BPS) yaitu lembaga pemerintah non kementerian yang bertanggungjawab langsung kepada presiden dengan salah satu tugasnya adalah menyediakan kebutuhan

data bagi pemerintah dan masyarakat. Data tersebut diperoleh dari kegiatan sensus/survey yang dilakukan oleh BPS atau dari departemen/lembaga pemerintah lainnya sebagai data sekunder. Dalam keseharian, pengertian istilah statistik dianggap sama dengan statistika, sehingga penggunaannya seringkali tertukar.

Statistik adalah teknik pengumpulan, pengolahan, penyajian serta penarikan kesimpulan dari suatu data yang digunakan dalam proses pengambilan keputusan

Statistika adalah suatu ilmu yang mempelajari teknik pengumpulan, pengolahan, penyajian serta penarikan kesimpulan dari suatu data yang digunakan dalam proses pengambilan keputusan

Secara umum statistika dibedakan menjadi dua yaitu **statistika deskriptif** dan **statistika inferensial**. Statistika deskriptif adalah ilmu yang mempelajari teknik pengumpulan, pengolahan dan penyajian data ke dalam bentuk grafik, bagan serta tabel sehingga data lebih mudah dipahami. Sedangkan statistika inferensial adalah ilmu yang mempelajari penarikan kesimpulan dari suatu data sehingga dapat digunakan sebagai pertimbangan dalam pengambilan keputusan.

2. Populasi dan Sampel

Dalam masa pemilihan presiden Indonesia 2019, suatu lembaga survey A akan melakukan survey di seluruh provinsi untuk mengetahui sejauh minat masyarakat untuk berpartisipasi dalam pesta demokrasi yang digelar lima tahun sekali tersebut dan bagaimana peta kekuatan masing-masing calon presiden. Mengingat dana yang terbatas, lembaga survey A tidak jadi mengambil data populasi di semua propinsi namun membatasi hanya mengambil data sampel di beberapa kota di pulau Jawa. Ilustrasi di atas mengenalkan tentang terminologi populasi dan sampel.

Populasi adalah kumpulan dari individu atau item yang mungkin diperoleh dari suatu pengamatan atau percobaan

Sampel adalah bagian dari populasi yang diambil untuk keperluan analisis data

3. Teknik Sampling

Selain dengan cara sensus, untuk mengumpulkan data dapat dilakukan dengan cara sampling dan percobaan. Dalam sub bab ini akan dipelajari bagaimana cara sampling yang benar melalui beberapa teknik sampling. Yang pertama **simple random sampling** adalah suatu prosedur sampling di mana setiap sampel dengan ukuran tertentu mempunyai kesempatan yang sama untuk dipilih. Sedangkan simple random sample adalah suatu sampel yang diperoleh dengan teknik simple random sampling. Contoh dari teknik simple random sampling adalah random number table yaitu mengeluarkan kertas yang berisi angka secara acak dari suatu kotak undian. Yang kedua **systematic random sampling** adalah suatu teknik random sampling dengan cara membagi ukuran populasi dengan ukuran sampel dan hasilnya dibulatkan ke bawah mendekati m. Menggunakan metode random number table untuk memperoleh suatu angka, k , yang berada antara 1 dan m, kemudian pilih sampel yang mempunyai angka k, k+m, k+2m, ... Yang ketiga **cluster sampling** adalah suatu teknik sampling dengan cara membagi populasi menjadi beberapa kelompok, gunakan simpel random sampel untuk mendapatkan sampel dari setiap kelompoknya. Sampel dari setiap grup inilah yang akan menjadi sampel.

4. Variabel dan Data

Dalam suatu penelitian mengenai tingkat populasi di kota besar, peneliti menggunakan variabel tingkat polusi untuk mengetahui seberapa tinggi tingkat polusi di kota tersebut. Jadi peneliti perlu paham apa arti variabel itu sendiri.

variabel adalah suatu karakteristik atau atribut yang dapat diasumsikan dengan nilai yang berbeda

Data tidak dipisahkan dalam kehidupan sehari-hari. Sebagai contoh ketika akan mendaftar kuliah jurusan matematika, kita memerlukan data mengenai universitas mana saja yang membuka jurusan matematika, berapa banyak pendaftar di jurusan tersebut dari tahun ke tahun dan seberapa tinggi tingkat persaingan masuk di jurusan tersebut.

Data adalah nilai dari suatu variabel yang diperoleh dari suatu percobaan atau pengamatan

Selain itu terdapat jenis data berdasarkan cara pengambilannya, yaitu **data primer** jika data diperoleh dari hasil percobaan atau survey yang dilakukan oleh peneliti sendiri, **data sekunder** jika data diperoleh dari data yang sudah dipublikasikan oleh pihak lain.

Berdasarkan jenisnya, data dapat dibedakan menjadi dua yaitu data **data kuantitatif** jika berupa angka, misalnya penghasilan karyawan dan jumlah penduduk di suatu kota. **data kualitatif** jika berupa kategori, misalnya data tentang jenis warna pelangi dan jenis kelamin.

Berdasarkan cara pengambilannya, data dapat dibedakan menjadi dua yaitu data **data diskrit** dan **data kontinu**. Data diskrit adalah data yang diperoleh dari hasil pencacahan, contohnya adalah banyaknya siswa yang datang tepat waktu, banyaknya buah manggis yang dijual di supermarket A. Data kontinu adalah data yang diperoleh dari hasil pengukuran, contohnya adalah tinggi badan siswa TK Suka CIta, kadar leukosit penderita demam berdarah di kota Nganjuk.

Selain itu terdapat jenis data berdasarkan cara pengambilannya, yaitu **data primer** jika data diperoleh dari hasil percobaan atau survey yang dilakukan oleh peneliti sendiri, **data sekunder** jika data diperoleh dari data yang sudah dipublikasikan oleh pihak lain.

Tabel 1.1: Contoh Data Kualitatif dan Kuantitatif

Jenis Data		Pertanyaan yang diajukan	Respons
Kualitatif		Status perkawinan?	☐ belum kawin ☐ kawin ☐ duda/janda
Kuantitatif	Diskret	Berapa jumlah balita di rumah Anda?	... anak
	Kontinu	Berapa usia Anda?	... tahun ... bulan

Tabel 1.2: Contoh Data Nominal dan Ordinal

skala pengukuran	data kualitatif	kategori	
nominal	jenis kelamin	pria	wanita
	nikah	ya	tidak
	warna favorit	merah	kuning
		hijau	biru
ordinal	pendidikan terakhir	sd	smp
		sma	s1
	jabatan fungsional dosen	asisten ahli	lektor
		lektor kepala	guru besar

5. Rangkuman

Pada sub bab rangkuman ada beberapa hal yang perlu diingat oleh pembaca, hal hal tersebut adalah sebagai berikut.

Statistik adalah data sedangkan statistika adalah suatu ilmu yang mempelajari teknik pengumpulan, pengolahan, penyajian serta penarikan kesimpulan dari suatu data yang digunakan dalam proses pengambilan keputusan.

variabel adalah suatu karakteristik atau atribut dan data adalah nilai dari suatu variabel yang diperoleh dari suatu percobaan atau pengamatan.

C. Perlatihan

I. Tentukan soal di bawah ini merupakan statistika deskriptif atau statistika inferensial

(1) Rata-rata harga tayang iklan selama 30 detik di chanel youtube Atta Halilintar adalah Rp.100.000,00.

(2) Dari hasil survey yang dilakukan kepada 234 orang pelanggan PT.KAI, disimpulkan bahwa separuh responden menyatakan ketidakpuasan menggunakan sarana toilet kereta api ekonomi

II. Tentukan soal di bawah ini merupakan data diskrit atau kontinu

(1) Banyaknya jam yang digunakan oleh anak balita menonton film kartun

(2) Berat kotor beras premium yang dijual di supermarketi

D. Bahan Bacaan

Bluman, A. G. 2018. *Elementary Statistics: A Step By Step Approach (10th Edition)*. New York: McGraw-Hill Education.

Walpole, M. 2013. *Probability and Statistics for Engineers and Scientist: Global Edition (9th Edition)*. Mumbai: Pearson.

Weiss, N. A. 2017. *Introductory Statistics: Global Edition (10th Edition)*. Harlow: Pearson.

BAB II
UKURAN PEMUSATAN DAN PENYEBARAN DATA

A. Pendahuluan

1. Deskripsi Singkat

Dalam bab ini akan dibahas pengertian tentang ukuran ukuran pemusatan dan penyebaran data.

2. Tujuan Pembelajaran

a. Menjelaskan pengetahuan tentang ukuran pemusatan data
b. Menjelaskan pengetahuan tentang variabel ukuran penyebaran data
c. Menjelaskan pengetahuan tentang formula untuk ukuran pemusatan dan penyebaran data

B. Materi

1. Ukuran Pemusatan Data

Ukuran deskriptif yang mengindentifikasikan di mana letak pusat atau kecenderungan nilai dari suatu data set disebut ukuran pusat kecenderungan yang lebih dikenal dengan ukuran pemusatan data. Pada sub bab ini akan didiskusikan tiga ukuran pemusatan data yang sering digunakan dalam kehidupa sehari-hari, yaitu mean, median dan modus.

Mean Populasi atau rata-rata dari suatu populasi data adalah jumlahan dari seluruh anggota populasi dibagi dengan banyaknya anggota populasi.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1}$$

Mean Sampel atau rata-rata dari suatu sampel data adalah

jumlahan dari sampel dibagi dengan banyaknya sampel.

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n} \tag{2.2}$$

Contoh 2.1. Diberikan data nilai mata kuliah Metode Statistika mahasiswa kelas 2019 A sebagai berikut: 57, 86, 79, 45, 89, 72, 94, 68, 44, 23, 97, 65, 49, 83, 79 Tentukan rata-rata nilai di atas.

Penyelesaian

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n} = \frac{57+86+79+45+89+72+94+68+44+23+97+65+49+83+79}{15} = 68,67$$

jadi rata-rata dari data nilai mata kuliah Metode Statistika mahasiswa kelas 2019 A adalah 68,67. ◀

Contoh 2.2. Seorang pedagang kelontong ingin mengetahui rata-rata penjualan dalam satu minggu, data yang dia punya adalah sebagai berikut: 175000, 325700, 55500, 895600, 485000, 274500, 624500 Tentukan rata-rata penjualan pedagang tersebut.

Penyelesaian

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n} = \frac{175000+325700+55500+895600+485000+274500+624500}{7} = 405114,29$$

jadi rata-rata penjualan pedagang tersebut dalam satu minggu adalah 405114,29. ◀

Modus dari suatu data set adalah data yang memiliki frekuensi terjadinya paling tinggi.

Contoh 2.3. Hasil pengukuran tinggi badan (dalam cm) 20 siswa kelas 7 suatu SMP swasta di Surabaya adalah sebagai berikut: 155, 147, 160, 132, 153, 147, 128, 159, 147, 153, 136, 168, 129, 118, 147, 150, 152, 140, 161, 132. Tentukan modus dari data di atas.

Penyelesaian Modus dari data di atas adalah 147 karena nilai tersebut yang paling sering terjadi/muncul. ◀

Contoh 2.4. Berat badan bayi (dalam kg) yang mengikuti program posyandu desa Sukamaju adalah sebagai berikut: 3.3, 3, 2.7, 4.2, 3.1, 3.9, 3,2, 5, 4.8, 3.7 Tentukan modus dari data di atas.

Penyelesaian Data di atas tidak mempunyai modus karena masing-masing nilai hanya muncul sekali. ◀

Median suatu data set adalah nilai yang terletak di tengah setelah data telah diurutkan dari kecil ke besar. Jika banyaknya pengamatan adlah ganjil maka median data akan terletak tepat pada tengah urutan. Jika banyaknya pengamatan adalah genap, maka median data terletak di rata-rata nilai tengah urutan.

Contoh 2.5. Usia sebelas lansia yan ikut posyandu lansia di desa Sukamaju adalah sebagai berikut: 59, 51, 60, 66, 63, 83, 68, 67, 77, 70, 71 Tentukan median dari data di atas.

Penyelesaian Untuk menyelesaikan soal diatas maka langkah langkah yang diperlukan ada di Tabel 2.1. ◀

Langkah 1	urutkan data mulai dari yang terkecil: $51, 59, 60, 63, 66, 67, 68, 70, 71, 77, 83$
Langkah 2	tentukan banyaknya data: n=11
Langkah 3	tentukan letak median $(\frac{n+1}{2})$
Langkah 4	median dari data di atas adalah 67 yang terletak pada urutan ke enam.

Tabel 2.1: Penyelesaian untuk contoh 2.5

Contoh 2.6. Usia sepuluh lansia yang ikut posyandu lansia di desa Sukamaju adalah sebagai berikut: 59, 51, 60, 66, 63, 83, 68, 67, 77, 70 Tentukan median dari data di atas.

Penyelesaian Untuk menyelesaikan soal diatas maka langkah langkah yang diperlukan ada di Tabel 2.2. ◀

2. Ukuran Penyebaran Data

Untuk dapa mendeskripsikan suatu data secara rinci, selain diperlukan pengetahuan mengetahui ukuran pemusatan data juga diperlukan pengetahuan mengenai ukruan penyebaran data. Ukuran penyebaran data yang dibahas dalam sub bab ini antar lain range, variansi dan standard deviasi.

Range suatu data set adalah selisih antara nilai pengamat terbesar dengan nilai pengamatan terkecil dan contohnya bisa dilihat di 2.7.

Langkah 1	urutkan data mulai dari yang terkecil: $51, 59, 60, 63, 66, 67, 68, 70, 77, 83$
Langkah 2	tentukan banyaknya data: n=11
Langkah 3	tentukan letak median $(\frac{n+1}{2})$
Langkah 4	median dari data di atas adalah 66,5 yang terletak di antara urutan ke lima dan enam.

Tabel 2.2: Penyelesaian untuk contoh 2.6

Contoh 2.7. Usia sepuluh lansia yang ikut posyandu lansia di desa Sukamaju adalah sebagai berikut: 59, 51, 60, 66, 63, 83, 68, $67, 77, 70$ Tentukan range dari data di atas.

Penyelesaian Untuk menjawab contoh 2.7, maka perlu dilakukan tahap tahapan pada Tabel 2.3. ◀

Langkah 1	urutkan data mulai dari yang terkecil: $51, 59, 60, 63, 66, 67, 68, 70, 77, 83$
Langkah 2	kurangkan nilai terbesar dengan nilai terkecil
Langkah 3	range dari data di atas adalah 32

Tabel 2.3: Penyelesaian contoh 2.7

Selain range seperti yang disebut diatas, maka jika kita menggunakan mean sebagai pusat data, untuk penyebaran datanya kita bisa menggunakan varians. Sedangkan definisinya adalah sebagai berikut :

Apabila kita memilih mean sebagai ukuran pemusatan data maka ukuran penyebarannya adalah **varians**.
variansi Populasi adalah seberapa jauh perbedaan atau jarak setiap nilai observasi dalam suatu populasi dari rata-ratanya.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)}{n} \tag{2.3}$$

Standard Deviasi populasi adalah akar kuadrat dari variansi populasi

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}} \tag{2.4}$$

Variansi Sampel adalah seberapa jauh perbedaan atau jarak setiap nilai observasi sampel dari rata-ratanya.

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{2.5}$$

Standard deviasi Sampel adalah akar kuadrat dari variansi sampel.

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \tag{2.6}$$

Sebagai illustrasi untuk mencari varians adalah sebagai berikut.

Contoh 2.8. Diketahui data lamanya bertahan hidup dari sekelompok anak burung adalah sebagai berikut: 10, 60, 50, 30, 40, 20 Tentukan variansi dan standard deviasi populasi dari data di atas.

Penyelesaian Langkah langkah untuk menyelesaikan contoh 2.8 ada pada Tabel 2.4. ◀

Langkah 1	hitunglah rata-rata dari data $$\bar{x} = \frac{\sum X}{N} = \frac{10+60+50+40+30+20}{6} = \frac{210}{6} = 35$$
Langkah 2	kurangkan setiap nilai dengan rata-rata data $$x - \bar{x}$$
Langkah 3	kuadratkan setiap hasil dari langkah 2 $$(x - \bar{x})^2$$
Langkah 4	jumlahkan hasil dari langkah 3 $$\sum (x - \bar{x})^2 = 625 + 625 + 225 + 25 + 25 + 225 = 1750$$
Langkah 5	hasil langkah 4 dibagi dengan banyaknya pengamatan (n) $$\frac{\sum (x - \bar{x})^2}{n} = \frac{1750}{6} = 291,7$$
Langkah 6	Sehingga varians dari populasinya adalah 291,7
Langkah 7	menghitung akar kuadrat dari hasil langkah 5 $$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{291,7} = 17,1$$
Langkah 8	Standard deviasi dari data di atas adalah 17,1

Tabel 2.4: Penyelesaian contoh 2.8

Apabila kita memilih ukuran pemusatannya adalah median, maka ukuran penyebarannya adalah *interquartile range* (IR).

$$IR = Q_3 - Q_1 \tag{2.7}$$

Untuk contoh dengan menggunakan Sofware R adalah sebagai berikut :

Contoh 2.9. Usia sepuluh lansia yang ikut posyandu lansia di desa Sukamaju adalah sebagai berikut: 59, 51, 60, 66, 63, 83, 68, 67, 77, 70 Tentukan mean, median, varians, dan standar deviasi dari data di atas.

Penyelesaian R script dalam R studio adalah sebagai berikut :

```
data <-c(59, 51, 60, 66, 63, 83,  68, 67, 77, 70)
summary(data) # untuk menghitung lima ukuran
var(data) # menghitung varians data
sd(data) # menghitung standard deviasi dari data.
```

Sedangkan output dari script diatas adalah

```
> summary(data)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
51.00   60.75   66.50   66.40   69.50   83.00
> var(data)
[1] 83.15556
> sd(data)
[1] 9.118967
```

◀

3. Rangkuman

(1). Mean Populasi atau rata-rata dari suatu populasi data adalah jumlahan dari seluruh anggota populasi dibagi dengan banyaknya anggota populasi

(2). Mean Sampel atau rata-rata dari suatu sampel data adalah jumlahan dari sampel dibagi dengan banyaknya sampel

(3). Modus dari suatu data set adalah data yang memiliki frekuensi terjadinya paling tinggi

(4). Median suatu data set adalah nilai yang terletak di tengah setelah data telah diurutkan dari kecil ke besar

(5). Jika banyaknya pengamatan adlah ganjil maka median data akan terletak tepat pada tengah urutan. Jika banyaknya pengamatan adalah genap, maka median data terletak di rata-rata nilai tengah urutan
(6). Range suatu data set adalah selisih antara nilai pengamat terbesar dengan nilai pengamatan terkecil
(7). Variansi Populasi adalah seberapa jauh perbedaan atau jarak setiap nilai observasi dalam suatu populasi dari rata-ratanya
(8). Variansi Sampel adalah seberapa jauh perbedaan atau jarak setiap nilai observasi sampel dari rata-ratanya
(9). Standard Deviasi adalah akar kuadrat dari variansi

C. Perlatihan

Diketahui data adalah sebagai berikut :
$44, 12, 22, 31, 26, 22, 30, 26, 18, 28, 12, 40, 17, 13, 14, 17, 25, 29,$ dan $15, 30, 10, 28, 16, 33, 24, 20, 29, 34, 23, 13$

a. Buatlah histogram !
b. Buatlah boxplot !
c. Hitung ukuran pemusatan yang sesuai !
d. Hitung ukuran penyebaran yang sesuai !.

D. Buku Bacaan

Bluman, A. G. 2018. *Elementary Statistics: A Step By Step Approach (10th Edition)*. New York: McGraw-Hill Education.

Walpole, M. 2013. *Probability and Statistics for Engineers and Scientist: Global Edition (9th Edition)*. Mumbai: Pearson.

Weiss, N. A. 2017. *Introductory Statistics: Global Edition (10th Edition)*. Harlow: Pearson.

BAB III
VISUALISASI DATA

A. Pendahuluan

1. Deskripsi Singkat

Dalam bab ini akan dibahas pengertian tentang memvisualisasi data

2. Tujuan Pembelajaran

a. Menjelaskan pengetahuan tentang tabe distribusi frekuensi
b. Menjelaskan pengetahuan tentang histogram
c. Menjelaskan pengetahuan tentang boxplot

B. Materi

1. Histogram

Data yang telah kita peroleh/kumpulkan, data dapat disajikan dalam bentuk grafik secara statistika yang bertujuan untuk memudahkan pembaca membaca data. Salah satu bentuk grafik yang dibahas dalam bab ini adalah histogram. Dalam menggambar histogram, diperlukan langkah pertama yang dilakukan adalah membuat tabel distribusi frekuensi dengan tahapan sebagai pada Tabel 3.1.

Setelah mendapatkan tabel frekuensi maka langkah selanjutnya ada pada Tabel 3.2.

Contoh 3.10. 12, 14, 14, 14, 16, 18, 20, 20, 21, 23, 27, 27, 27, 29, 31, 31, 32, 32, 34, 36, 40, 40, 40, 40, 40, 42, 51, 56, 60, 65, buatlah histogram dari data tersebut!.

Penyelesaian Untuk menyelesaikan soal diatas maka hal pertama yang perlu dilakukan adalah membuat tabel frekuensinya dengan langkah langkah pada Tabel 3.3.

Sedangkan untuk langkah selanjutnya adalah kita pilih 12 sebagai nilai terendah kemudian kita jadika nilai tersebut sebagai

Langkah-Langkah Membuat tabel frekuensi	
Langkah 1	etntukan banyaknya kelas
Langkah 2	tentukan lebar kelas yang dapat diperoleh dari nilai jangkauan yang dibagi oleh banyaknya kelas
Langkah 3	bulatkan lebar kelas yang telah didapat
Langkah 4	tulis nilai paling kecil sebagai limit batas bawah kelas
Langkah 5	tambahkan nilai pada tahap sebelumnya dengan lebar kelas untuk mendapatkan limit bawah kelas berikutnya
Langkah 6	tulis batas atas limit kelas

Tabel 3.1: Langkah-Langkah Membuat tabel frekuensi

Langkah-Langkah Membuat Histogram	
Langkah 1	buat koordinat kartesius dengan sumbu x mewakili banyak kelas dalam distribusi frekuensi, sedangkan sumbu y mewakili frekuensi setiap kelasnya
Langkah 2	gambar persegi panjang untuk setiap kelas yang ada dalam tabel distribusi frekuensi

Tabel 3.2: Langkah-Langkah Membuat Histogram

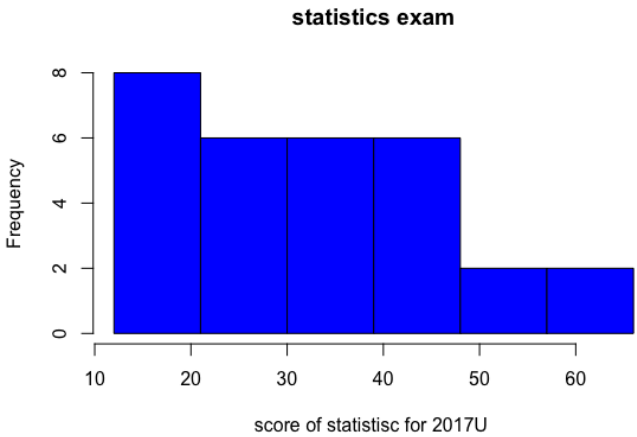
Langkah-Langkah Membuat tabel frekuensi	
Langkah 1	misal : banyaknya kelas adalah $n = 6$
Langkah 2	lebar kelas $= \frac{jangkauan}{n} = \frac{max-min}{6} = \frac{65-12}{6} = 8,8$
Langkah 3	bulatkan lebar kelas sehingga menjadi 9
Langkah 4	tulis nilai paling kecil sebagai limit batas bawah kelas.

Tabel 3.3: Penyelesaian untuk contoh 3.10

batas bawah limit kelas. Untuk mendapatkan batas bawah limit kelas, bisa kita tambahkan nilai 12 dengan lebar kelasnya, yaitu 9 sehingga didapatkan nilai 21. Selanjutnya kita tulis batas atas limit kelas, yaitu 20. Kita ulangi prosedur tersebut sampai dengan jumlah kelas yang telah ditentukan pada tahap awal. Tabel frekuensi dari contoh 3.10 terdapat pada Tabel 3.4.

Interval	Frequency
12-20	8
21-29	6
30-38	6
39-47	6
48-56	2
57-65	2

Tabel 3.4: Tabel frekuensi untuk contoh 3.10



Gambar 3.1: Histogram untuk contoh 3.10

Setelah itu baru kita bisa menggambar histogram dari tabel frekuensi tadi dengan prosedur pada Tabel 3.2 dan didapat seperti pada gambar Gambar 3.1. ◀

Untuk R scripts dari histogram untuk data tersebut adalah sebagai berikut :

```
data<-c(12, 14, 14, 14, 16, 18, 20, 20, 23, 23, 27, 27, 27,
29, 31, 31, 32, 32, 34, 36, 40,
40, 40, 40, 40, 42, 51, 56, 60, 65)
# histogram yang kasar
hist(data)
```

```
### menggambar histogram dengan batas bawah 12
###  dan lebar kelasnya adalah 9
### membuat kelas dengan memasukkan lebar kelas
bins <- seq(12,66,by=9)
### memasukkan nilai yang cocok kedalam kelas
interval<-cut(data,bins)
table(interval) ## table frekuensi
transform(table(interval)) ### tampilan table frekuensi

# memberikan judul pada histogram
hist(data,breaks=bins,main="statistics exam",
xlab="score of statistisc for 2017U ", col="blue")
```

Contoh 3.11. Tabel di bawah ini memberi informasi mengenai suhu di 50 daerah. Data telah dibuat dalam tabel distribusi frekuensi sebagai berikut: Buatlah histogram dari data di atas.

Interval	Frequency
99,5 - 104,5	2
104,5-109,5	8
109,5-114,5	18
114,5-1119,5	13
119,5-124,5	7
124,5-129,5	1
129,5-134,5	1

Tabel 3.5: Tabel suhu di 50 daerah

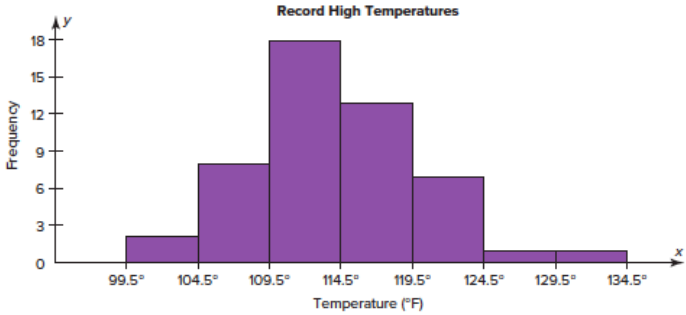
Penyelesaian Histogramnya terdapat pada Gambar 3.2. ◀

2. Box-plot

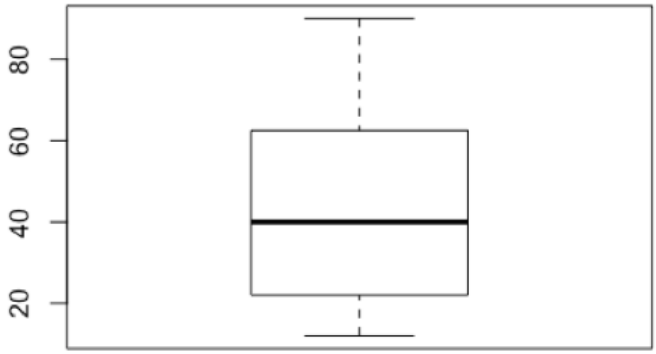
Visualisasi data yang kedua adalah Boxplot yang digunakan untuk mendeteksi secara visual adanya outlier dalam data. Dan untuk menggambar box-nya diperlukan median, Q_1 dan Q_2.

Contoh 3.12. Diketahui data daya tahan lampu (dalam bulan) sebagai berikut: 41, 36, 12, 18, 23, 45, 76, 82, 40, 15, 45, 80, 24, 21, 49, 79, 12, 29, 90, 40. Buatlah boxplot dari data tersebut.

Penyelesaian Boxplot yang dibentuk terdapat pada Gambar 3.3.



Gambar 3.2: Histogram untuk contoh 3.11



Gambar 3.3: Boxplot untuk contoh 3.12

Langkah-Langkah Membuat Boxplot	
Langkah 1	hitung semua quartile data
Langkah 2	buat garis berskala dan diberi tanda untuk setiap quartile
Langkah 3	hitung jangkauan antar kuartile, range dan tentukan posisi batas outlier
Langkah 4	posisi outlier atas diperoleh dengan menambahkan jangkauan antar quartile dengan quartile ketiga
Langkah 5	tentukan outlier dan hapus batas outlier

◀

Apabila ingin diselesaikan dengan Rstudio maka scriptnya adalah sebagai berikut :

```
data<-c(41, 36, 12, 18, 23, 45, 76, 82, 40, 15, 45, 80,
24, 21, 49, 79, 12, 29, 90, 40)
boxplot(data)
```

3. Rangkuman

Histogram dan boxplot merupakan bentuk penyajian data agar lebih mudah dibaca dan dipahami

C. Perlatihan

Buatlah histogram dan boxplot dari setiap soal di bawah ini:

(1). Hasil data dari penelitian kadar eritrosit para penderita demam berdarah adalah sebagai berikut: 66, 52, 47, 43, 69, 59, 56, 39, 57, 56, 61, 50, 45, 49.

(2). Data banyaknya pendatar jurusan matematika universitas X selama 20 tahun terakhir adalah sebagai berikut: 85, 90, 65, 87, 90, 35, 45, 76, 89, 23, 34, 23, 57, 59, 89, 70, 77, 80, 91, 50.

(3). Data modal yang diinvestasikan dalam bidang telekomunikasi (dalam juta rupiah) adalah sebagai berikut:3344, 434, 3291, 4617, 5875, 149, 18000, 3818, 749, 7091, 179, 3536, 4163, 3615, 9324, 4969, 177, 8297, 515, 2455, 11191, 6593, 345, 192, 1528, 8083, 1936.

(4) Diketahui data adalah sebagai berikut : $44, 12, 22, 31, 26, 22, 30, 26, 18, 28, 12, 40, 17, 13, 14, 17, 25, 29$, dan $15, 30, 10, 28, 16, 33, 24, 20, 29, 34, 23, 13$

a. Buatlah boxplot !

b. Hitung ukuran pemusatan yang sesuai !

c. Hitung ukuran penyebaran yang sesuai !.

D. Buku Bacaan

Bluman, A. G. 2018. *Elementary Statistics: A Step By Step Approach (10th Edition)*. New York: McGraw-Hill Education.

Walpole, M. 2013. *Probability and Statistics for Engineers and Scientist: Global Edition (9th Edition)*. Mumbai: Pearson.

Weiss, N. A. 2017. *Introductory Statistics: Global Edition (10th Edition)*. Harlow: Pearson.

BAB IV
RUANG SAMPEL, PELUANG DAN DISTRIBUSI PELUANG

A. Pendahuluan

1. Deskripsi Singkat

Pada bab sebelumnya, statistika deskriptif yang meliputi ukuran pemusatan, ukuran penyebaran serta visualisasi data telah dibahas. Sehingga pada bab ini akan disajikan beberapa pengetahuan dasar untuk menganalisis data dengan menggunakan statistika inferensial. Konsep dasar tersebut terdiri atas ruang sampel, peluang dan beberapa distribusi peluang

2. Tujuan Pembelajaran

a. Menjelaskan pengetahuan tentang ruang sampel
b. Menjelaskan pengetahuan tentang peluang
c. Menjelaskan pengetahuan tentang variabel acak dan distribusi peluang

B. Materi

1. Ruang Sampel

Didalam kehidupan sehari hari dapat kita temukan peristiwa peristiwa yang merupakan percobaan statistik. A simple example of a statistical experiment is the tossing of a coin. In this experiment, there are only two possible outcomes, heads or tails. Salah satu contoh yang paling popular adalah adalah ketika seseorang meleparkan sebuah dadu yang bermata sebanyak enam.

Ruang sampel adalah sebuah himpunan dari seluruh kemungkinan hasil yang muncul dari suatu percobaan statistika dan disimbolkan dengan S
Titik sampel adalah setiap anggota dari ruang sample S
Ukuran sampel adalah jumlah total banyaknya titik sampel dari suatu ruang sampel

Apabila orang tersebut fokus pada mata dadu yang muncul ketika dadu tersebut dilempar, maka himpunan segala kemungkinan mata dadu yang muncul ketika melepar dadu disebut sebagai ruang sampel. Sedangkan simbol dari ruang sampel adalah S.
Oleh karena itu ruang sampel dari percobaan statistika diatas dapat ditulis sebagai berikut :

$$S = \{1, 2, 3, 4, 5, 6\}$$

dan setiap anggota dari ruang sampel S dapat ditulis sebagai berikut :

$$1, 2, 3, 4, 5$$

dan 6 dikatakan sebagai titik sampel. Banyaknya titik sampel dari melepar sebuah dadu diatas sebanyak 6 atau bisa dikatakan bahwa ukuran sampelnya adalah 6.

Pada percobaan statistika melepar sebuah dadu, terkadang ada hal yang menarik yang bukan sekedar ingin mengetahui mata dadu yang muncul setiap kali dilempar. Misalnya adalah munculnya mata dadu yang bisa dibagi 3 dimana peristiwa ini kita simbolkan dengan A. Ruang sampel dan titik sampel dari A adalah

$$A = \{3, 6\}$$

dan titik sampelnya adalah 3 dan 6.

Berdasarkan uraian diatas bisa diketahui bahwa sebenarnya anggota dari A adalah bagian dari S. Dengan kata lain bahwa A merupakan kejadian dari S dan Amerupakan himpuanan bagian dari S

Kejadian merupakan himpunan bagian dari ruang sampel S

Contoh 4.13. Apabila sebuah koin dilempar sebanyak tiga kali, tentukan : ruang sampel, titik sampel, suatu kejadian dan kejadian yang dibentuk oleh beberapa kejadian sebelumnya

Penyelesaian Koin pada dasarnya memiliki dua sisi, yaitu angka (A) dan gambar (G). Sehingga

- ruang sampelnya adalah

  $$S = \{AAA, AAG, AGA, AGG, GAA, GAG, GGA, GGG\}$$

- titik sampelnya adalah $AAA, AAG, AGA, AGG, GAA, GAG, GGA$ dan GGG dan jumlahnya adalah 8

- kejadian B adalah munculnya sisi angka A minimal 1. Sehingga ruang sampel B adalah

  $$B = \{AAA, AAG, AGA, AGG, GAA, GAG, GGA\}$$

  kejadian C adalah munculnya sisi gambar G minimal 2. Sehingga ruang sampel untuk C dalah

  $$C = \{AGG, GAA, GAG, GGA, GGG\}$$

- kejadian D adalah munculnya sisi angka G sebanyak 2. Ruang sampel untuk D adalah sebagai berikut :

  $$D = \{AGG, GAG, GGA\}.$$

  Dari uraian diatas, bahwa kejadian D beranggotakan sebagaian dari kejadian B dan C atau dengan kata lain $D = B \cap D$.

◀

Dari illustrasi contoh diatas, bahwa kejadian bisa didapatkan dari operasi antar kejadian. Operasi tersebut bisa $\cap, \cup$ ataupun

komplemen suatu kejadian. Untuk selanjutnya akan dibahas tentang bagaimana mendapatkan jumlah titik sampel dari suatu ruang sampel. Ada beberapa cara yang bisa digunakan,yaitu teknik perkalian, permutasi maupun kombinasi. Pada buku ini hanya akan fokus pada teknik perkalian.

Pada sub bab diatas, terdapat contoh melepar sebuah dadu yang bermata enam. Apabila dua dadu dilemparkan secara bersama sama, maka banyaknya titik sampel untuk dadu pertama (n_1) adalah 6 dan banyaknya titik sampel untuk dadu kedua adalah (n_2) adalah 6. Sehingga jumlah total titik sampel pada percobaan melempar dua dadu adalah $n_1 \times n_2 = 6 \times 6 = 36$. Metode tersebut dinamakan teknik perkalian.

2. Peluang dari suatu kejadian

Asal mula munculnya teori peluang berawal dari permainan judi. Teori peluang yang dikembangkan oleh para matematikawan berkembang pesat seiring dengan inginnya mengoptimalkan stategi untuk menang dalam suatu kompetisi. Kemudian seiring dengan perkembangan waktu, teori peluang meluas ke beberapa bidang seperti politik, bisnis, perkiraan cuaca maupun penelitian keilmuan.

Pada sub bab ini, kita fokus pada percobaan statistika dimana ruang sampelnya memuat elemen yang bisa dihitung (*finite*). Setiap anggota dalam ruang sampel tersebut tentukan probabilitasnya dimana total secara keseluruhan adalah 1. Jika kita mempunyai alasan kuat apabila suatu titik sample akan muncul pada percobaan statistika maka peluangnya mendekati 1. Sebaliknya, apabila kita menyakini bahwa titik sampel tersebut tidak muncul dalam suatu percobaan statistika maka peluangnya adalah 0.

Untuk mendapatkan peluang dari suatu kejadian A, kita akan menjumlahkan seluruh probabilitas yang telah ditetapkan pada titik sampel di A. Penjumlahan tersebut kita namakan sebagai peluang dari A dan dinotasikan sebagai $P(A)$.

Peluang dari kejadian A adalah total bobot dari seluruh titik sample didalam A. Maka

$$0 \leqslant P(A) \leqslant 1, \quad P(\varnothing) = 0, \quad \text{and} \quad P(S) = 1.$$

Sehingga, jika $A_1, A_2, \ldots$ adalah suatu deret dari kejadian yang *mutually exclusive*, maka

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$$

Contoh 4.14. Sebuah koin dilempar sebanyak dua kali. Berapa peluang minimum muncul satu gambar ?.

Penyelesaian Ruang sample dari percobaan tersebut adalah

$$S = \{GG, GA, AG, AA\}.$$

Jika kondisinya stabil, maka setiap kemungkinan diatas akan muncul. Oleh karena itu kita menetapkan suatu peluang ω untuk setiap titik sampel. Sehingga $4\omega = 1$ or $\omega = \frac{1}{4}$. Jika A adalah kejadian dari minimum gambar yang muncul, maka

$$A = \{GG, GA, AG\}$$

and

$$P(A) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}.$$

◀

Jika ruang sampel dari suatu percobaan memuat element sebanyak N yang kemungkinan akan muncul, maka kita menetapkan bahwa peluang yang sama sebesar $\frac{1}{N}$ untuk setiap titik titik N.

Peluang dari sebuah kejadian A yang memuat n dari N titik sampel adalah perbandingan antara banyaknya elemen pada A dan banyaknya element didalam S.

Jika suatu percobaan menghasilkan keluaran sebanyak N dan n merupakan hasil keluaran dari kejadian A, maka **peluang dari kejadian** A adalah

$$P(A) = \frac{n}{N}$$

Contoh 4.15. Pada kelas statistika di MIPA, mahasiswanya berasal dari 25 orang jurusan matematika, 10 orang jurusan fisika, 10 orang jurusan kimia dan 8 orang jurusan biologi. Jika seorang mahasiswa diambil secara acak untuk menjawab pertanyaan, tentukan peluang dari mahasiswa yang terpilih adalah
(a) berasal dari jurusan matematika
(b) berasal dari jurusan biologi atau jurusan fisika

Penyelesaian Kita misalkan bahwa M, F, K, B adalah mahasiswa dari jurusan matematika, fisika, kimia dan biologi secara berturut turut. Total mahasiswa dalam kelas adalah 53 dan semuanya memiliki kesamaan kesempatan untuk diseleksi.
(a) Jumlah mahasiswa matematika adalah 25 sedangkan total mahasiswa sekelas adalah 53. Maka peluang terpilihnya mahasiswa matematika (M) secara acak adalah sebagai berikut

$$P(M) = \frac{25}{53}.$$

(b) Jumlah mahasiswa jurusan biologi atau fisika adalah 18 orang dari 53 mahasiswa. Sehingga

$$P(B \cup F) = \frac{18}{53}.$$

◀

Teori peluang berkembang sangat pesat, mulai peluang bersyarat maupun *Bayes Rule's* yang merupakan fondasi dari ilmu bayesian.

3. Variabel Acak dan Distribusi Peluang

Pada dasarnya statistika adalah menganalisis data pada populasi dan mengeksplorasi tentang karakteristik dari populasi

tersebut. Suatu percobaan dalam menilai banyaknya komponen yang rusak adalah salah satu contoh dari percobaan statististika. Pada percobaan tersebut, hasil yang diperoleh akan ditandai dengan suatu deksripsi nilai. Misalnya, ruang sampel dari tiga komponen elektronik yang dites dapat ditulis sebagai berikut :

$$S = \{NNN, NND, NDN, DNN, NDD, DND, DDN, DDD\}$$

dimana N untuk simbol komponen elektronik yang tidak rusak dan D adalah simbol untuk komponen elektronik yang rusak.

Salah satu yang menjadi perhatian adalah banyaknya barang yang rusak. Sehinga setiap titik pada ruang sampel akan ditandai oleh suatu angka yaitu $0, 1, 2$ atau 3. Bilangan bilangan tersebut merupakan pembobotan acak yang bergantung pada hasil dari suatu percobaan. Angka angka tersebut bisa juga diasumsikan sebgai variabel acak X, yaitu banyaknya item yang rusak ketika tiga komponen elektronik dites.

Variabel acak adalah suatu fungsi yang menghubungkan antara bilangan riil dengan setiap elemen dari ruang sampel.

Huruf besar X merupakan variabel acak dan komponennya disimbolkan dengan huruf kecil x.

Apabila didefinisikan bahwa variabel acak X adalah sebagai berikut :

$$X = \begin{cases} 1, & \text{jika komponen rusak} \\ 0, & \text{jika komponen tidak rusak} \end{cases}$$

random variabel untuk 0 dan 1 dipilih untuk menerangkan dua kemungkinan yang muncul dari suatu percobaan yang kemudian disebut sebagai variabel acak *Bernoulli*.

Suatu variabel acak dikatakan variabel acak diskrit apabila seluruh kemungkinan yang muncul anggotanya dapat dihitung. Akan tetapi jika variabel random dimana anggota kemunginan yang dihasilkan berupa interval maka disebut sebagai variabel acak tidak diskrit atau kontinu.

Apabila kita tertarik pada peluang dari variabel acak X, banyaknya komponen elektronik yang cacat dari tiga komponen yang dites, maka akan diperoleh seperti Tabel 4.1 dibawah ini

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

Tabel 4.1: Peluang dari X, banyaknya komponen yang cacat dari tiga komponen elektronik yang ditest

Pasangan terurut antara variabel acak dan peluangnya atau dapat ditulis dengan $(x, P(X = x))$ disebut sebagai distribusi peluang. Karena element dari variabel acak X dapat dihitung maka termasuk variabel acak diskrit sehingga distribusi peluangnya juga bersifar diskrit. Salah satu contoh distribusi peluang yang terkenal yang sifatnya diskrit adalah distribusi *Bernoulli* yang telah disebutkan diatas.

Sedangkan distribusi peluang yang sifatnya tidak diskrit atau kontinu merupakan pasangan terurut $(x, P(X = x))$ dimana X adalah variabel acak yang bersifat interval. Distribusi peluang yang terkenal adalah normal distribusi yang akan kita bahas pada sub bab berikut ini.

4. Distribusi Normal

Distribusi peluang kontinu yang terkenal adalah distribusi normal dan kurvanya ada pada Gambar 4.1. Distribusi ini biasanya sangat cocok untuk menggambarkan fenomena pada alam, industri maupun penelitian. Contohnya dalah ukuran ukuran pada percobaan metrologi, data curah hujan yang secara umum bisa digambarkan dengan distribusi normal. Sedangkan fungsi peluang normal distribusi dapat ditulis sebagai berikut :

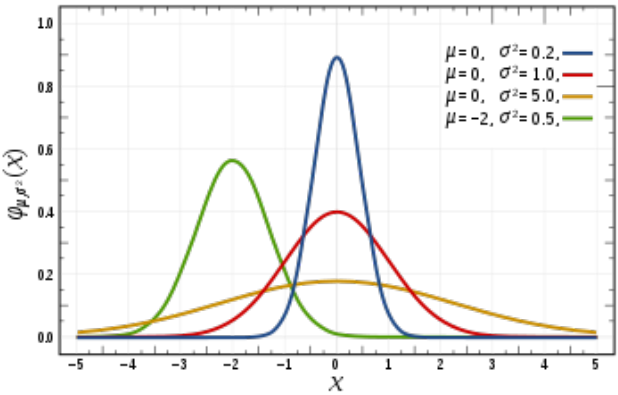
$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{4.1}$$

dimana $-\infty < x < \infty$ dan terdiri dari dua parameter yaitu : μ dan σ dan bersifat simetris.

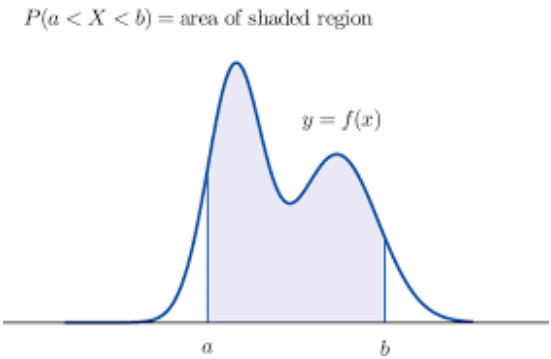
Apabila sekarang kita ingin mencari luasan dibawah kurva seperti Gambar 4.2. Maka perlu dilakukan transformasi dari x ke nilai z dengan rumusan sebagai berikut :

$$z = \frac{x - \mu}{\sigma}$$

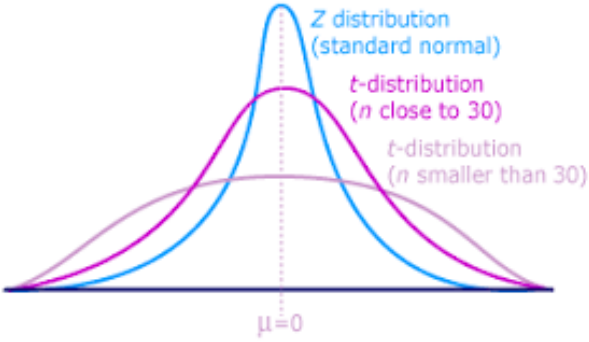
sehingga distribusi peluangnya menjadi standard distribusi normal dengan $\mu = 0$ dan $\sigma = 1$. Daerah dibawah kurva



Gambar 4.1: Kurva distribusi normal.



Gambar 4.2: Daerah dibawah kurva



Gambar 4.3: Kurva dari Distribusi Student's t.

untuk standar distribusi normal tidak perlu lagi kita hitung manual melainkan bisa menggunakan tabel kurva standar normal distribusi yang ada di lampiran.

Apabila data tidak memiliki σ maka kita perlu untuk mengestimasi dengan s dan apabila kurang dari 30 maka distribusi bukan lagi normal akan tetapi berubah menjadi distribusi T seperti pada Gambar 4.3. Tranformasi variabel baru dapat menggunakan formulasi berikut :

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

dengan derajat kebebasan (df)=$n - 1$.

Contoh 4.16. Diberikan distribusi normal dengan $\mu = 50$ and $\sigma = 10$. Tentukan peluang dari X yang terletak diantara 45 dan 62 !.

Penyelesaian Langkah langkahnya adalah sebagai berikut :

- lakukan transformasi variabel dari X ke Z
  misal : $x_1 = 45$ maka $z_1 = \frac{45-50}{10} = -0.5$
  misal : $x_2 = 62$ maka $z_2 = \frac{62-50}{10} = 1.2$
- didapat $P(45 \leqslant x \leqslant 62) = P(-0.5 \leqslant z \leqslant 1.2)$
- hitung $P(-0.5 \leqslant z \leqslant 1.2)$ dengan menggunakan tabel standar distribusi normal pada lampiran dan didapat sebagai

berikut :
$P(-0.5 \leqslant z \leqslant 1.2) = P(z \leqslant 1.2) - P(z \leqslant -0.5) = 0,8849 - 0,3085 = 0,5764$.

◀

5. Rangkuman

Jika suatu percobaan menghasilkan keluaran sebanyak N dan n merupakan hasil keluaran dari kejadian A, maka **peluang dari kejadian** A adalah

$$P(A) = \frac{n}{N}.$$

Apabila data memiliki σ maka dapat mengikuti distribusi normal. Akan tetapi apabila tidak memiliki σ maka kita perlu untuk mengestimasi dengan s dan apabila kurang dari 30 maka distribusi bukan lagi normal akan tetapi berubah menjadi distribusi T dengan df=$n - 1$.

C. Perlatihan

Kerjakan soal-soal berikut ini untuk mempertajam pemahaman Saudara!

(1) Sarah akan merakit komputer sendiri. Kemudian dia pilih *chip* dari dua merek, empat merek *hard drive*, tiga merek *memory* dan aksesoris dari lima toko komputer yang berbeda

Berapa cara Sarah memesan komponen komputer tersebut?.

(2) Diketahui bahwa X mengikuti distribusi normal dengan $\mu = 300$ dan $\sigma = 50$

Tentukan peluang dari X yang lebih besar dari 362?.

D. Bahan Bacaan

Bluman, A. G. 2018. *Elementary Statistics: A Step By Step Approach (10th Edition)*. New York: McGraw-Hill Education.

Walpole, M. 2013. *Probability and Statistics for Engineers and Scientist: Global Edition (9th Edition)*. Mumbai: Pearson.

Weiss, N. A. 2017. *Introductory Statistics: Global Edition (10th Edition)*. Harlow: Pearson.

BAB V
UJI HIPOTESIS, UJI Z DAN UJI T

A. Pendahuluan

1. Deskripsi Singkat

Setelah dalam bab sebelum kita membahas tentang peluang dan normal distribusi, maka bab selanjutnya adalah tentang statistika inferensial. Pada bab ini akan disajkan tentang uji hipothesis, uji Z dan uji T.

2. Tujuan Pembelajaran

a. Menjelaskan pengetahuan tentang istilah istilah uji hipotesis dan prosedurnya
b. Menjelaskan pengetahunan tentang uji Z dan uji T satu sampel
c. Menjelaskan pengetahunan tentang uji Z dan uji T dua sampel

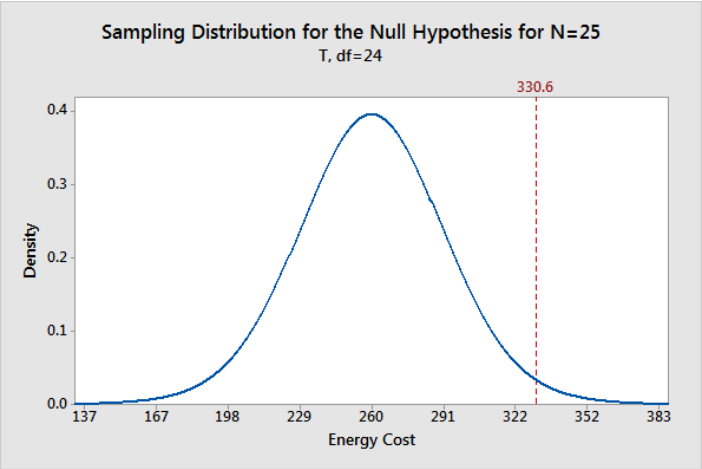
B. Materi

1. Uji Hipotesis

Seperti yang telah dibahas diawal bab bahwa statistika ada dua pendekatan yaitu statistika deskriptif dan statistika inferensial. Ada banyak hal yang terjadi dalam kehidupan sehari hari terkait pengambilan keputusan yang berkaitan dengan statistika inferensial. Misalnya, ketika kita diminta memilih dua produk yang berbeda merek namun penggunaannya sama atau memprediksi sesuatu yang akan datang berdasarkan data sebelumnya. statistika deskriptif tidak mampu lagi untuk menganalisis kebutuhan diatas. Untuk mengatasi hal tersebut, maka analisis statistika inferensial sangat dibutuhkan. Didalam analisis tersebut ada beberapa yang harus diketahui terlebih dahulu. Yang pertama adalah tentang hipotesis awal dan alternatif, *significant level* (α), *P-value* dan prosedur uji hipotesis.

Descriptive Statistics: Energy Cost

Variable	Total Count	Mean	SE Mean	StDev
Energy Cost	25	330.6	30.8	154.2



Gambar 5.1: Distribusi dari rata rata sampel

hipotesis awal dan alternatifnya

Seorang manager suatu sewa mobil ingin mengetahui apakah biaya perawatan mobil telah berubah dari tahun lalu, dimana perbulan rata menghabiskan 260 ribu rupiah. Untuk mengetahui hal tersebut maka manager tersebut mengambil sebanyak 25 mobil dan direkam biayanya selama setahun. Sampel tersebut diambil secara acakk dan rata-rata dari sampelnya adalah 330 ribu yang berbeda dari tahun sebelumnya, yaitu 260 ribu rupiah.

Dari illustrasi diatas, kita bisa menentukan dugaan awal yaitu rata rata pengeluaran perbulan untuk setiap mobilnya adalah sama dengan tahun lalu sebesar 26o ribu rupiah. Sehingga sebenarnya tujuan manager tadi adalah untuk mengetahui apakah rata rata sampel berbeda significant dengan dugaan awal.

Null hypothesis: The population mean equals the hypothesized mean (260).

Alternative hypothesis: The population mean differs from the hypothesized mean (260).

Untuk menjawab tujuan tersebut maka dbutuhkan suatu alat untuk mengambil keputusan yang disebut sebagai *significant level* dan *P-value*.

Seperti yang diterangkan diawal bahwa kita telah mendapatkan dugaan awal (*null hypothesis* (H_0)), yang didefinisikan dugaan yang mengandung sesuatu yang *null* (sama).

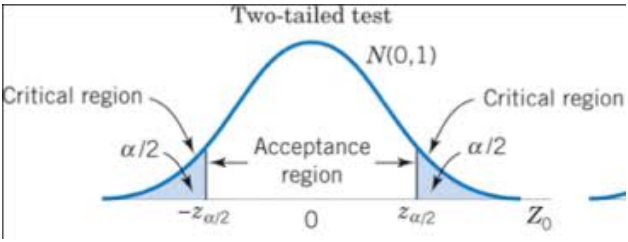
Sehingga kita bisa mendefinisikan bahwa
H_0 : rata rata pengeluaran untuk perawatan mobil secara keseluruhan (populasi) sama dengan tahun yang lalu yaitu 260 ribu rupiah
H_1 (alternatif dugaan) : rata rata pengeluaran untuk perawatan mobil secara keseluruhan (populasi) tidak sama dengan tahun yang lalu yaitu 260 ribu rupiah.

For example, a significance level of 0.05 indicates a 5 percent riskConcluding that a difference exists when there is no actual difference. Our sample mean which indicates it is statistically significant at the 0.05 level.

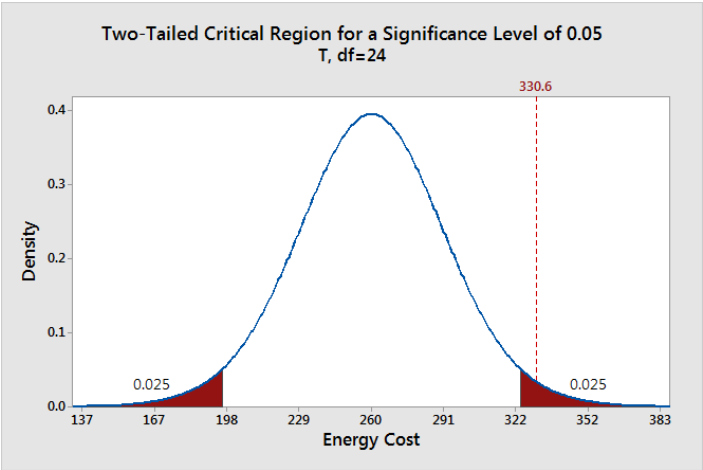
Dari dua dugaan diatas maka seorang peneliti hanya memiliki dua keputusan,yaitu gagal menolak H_0 atau menolah H_0. Seperti yang diungkapkan diatas, salah satu alat pengambil keputusan adalah *significant value* (α) yang didefiniskan sebagai peluang menolak H_0 ketika dugaan tersebut benar. Nilai dari α yang biasanya dipakai oleh peneliti ada 0.05 atau 0.01 bergantung jenis penelitiannya. Semakin beresiko penelitian tersebut, terutama berkaitan dengan makhluk hidup makan nilai α dipilih yang kecil yaitu 0.01. Sedangkan utnuk bidang sosial maka boleh memilih nilai α yang lebih besar.

Kita anggap dua ekor karena alternatif dugaan bisa mengandung dua makna yaitu rata pengeluaran lebih dari tahun lalu dan lebih kecil dari tahun lalu. Karena terdapat dua kemungkinan maka disebut dua ekor. Oleh karena itu nilai α dibagi dua yaitu setiap sisi bernilai sebesar $\frac{\alpha}{2}$. Dari nilai probabilitas $\frac{\alpha}{2}$ maka kita bisa mengetahui nilai batas pada sumbu x yang disebut sebagai nilai kritis $(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}})$.

Sehingga apabila nilai dari data terletak di daerah yang diarsir baik daerah sebelah kiri maupun kanan maka bisa ambil keputusan bahwa H_0 ditolak yang artinya adalah peneliti



Gambar 5.2: Daerah diterimanya dugaan awal (H_0)dari dua ekor (*two tails*)



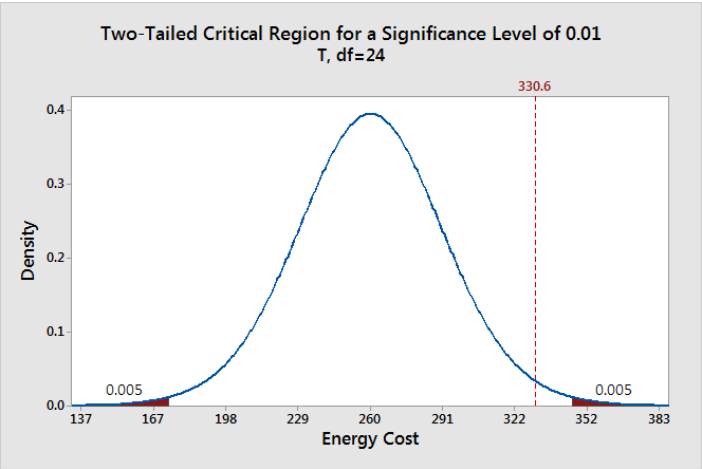
Gambar 5.3: Daaerah kritis *two tails* dengan *significant level*= 0.05

menerima alternatif dugaan. Sebaliknya jika nilai dari data terletak didaerah penerimaan maka keputusannya adalah gagal menolak H_0.

Kita kembali lagi ke ilustrasi dimana kita memilih $\alpha = 0.05$ maka hasil yang didapat dari data observasi atau sampel terletak didaerah kritis yang artinya adalah manager harus menolak H_0. Sehingga dapat disimpulkan bahwa pengeluaran untuk berbulan tidak sama dengan tahun yang lalu.

Apabila *significant level* kita ambil 0.01 maka hasil akan dapat dilihat seperti pada Gambar 5.4.

Dari Gambar 5.4 dapat diketahui bahwa hasil perhitungan



Gambar 5.4: Daaerah kritis *two tails* dengan *significant level*= 0.01

dengan menggunakan data observasi terletak di daerah penerimaan sehingga dapat disimpulkan bahwa ada fakta yang kuat untuk gagal menolak H_0. Dengan kata lain bahwa pengeluaran rata rata untuk perawatan mobil perbulan sama dengan tahun lalu yaitu sebesar 260 ribu rupiah.

Dari illustrasi diatas, dapat diketahui bahwa penentuan nilai *significant level* (α) sangat menentukan kesimpulan. Seperti contoh diatas , ketika menggunakan α= 0.05 kesimpulannya adalah tolak H_0 dan ketika α = 0.01 maka kesimpulannya menjadi sebaliknya yaitu gagal menolak H_0. Oleh karena itu tahapan penentuan *significant level* ini menjadi tahap yang sangat penting sebelum dilakukan analisis lebih lanjut.

Selain dengan *significant level*, kita bisa menggunakan *P-value*.Secara definisi *P-values* bertujuan untuk megukur seberapa besar data sampel mendukung bahwa H_0. Sehingga apabila nilai *P-value* tinggi maka ada kemungkinan bahwa data mendukung bahwa H_0 benar. Sebaliknya jika *P-value* rendah maka ada kemungkinan bahwa data tidak mendukung bahwa H_0 benar. Sehingga *P-value* rendah (kurang dari α) memberikan indikasi jika data sampel memberikan suatu dukungan yang kuat untuk menolak H_0 yang berlaku untuk seluruh populasi.

Untuk penerapan dari *P-value* akan ada di uji Z maupun T pada sub bab berikutnya.

Dari uraian diatas,dapat disimpulkan bahwa langkah langkah untuk melakukan analisis inferensial adalah sebagai berikut :

Langkah-langkah uji hipotesis	
Langkah 1	Menentukan dugaan awal dan alternatifnya.
Langkah 2	Menentukan *significant level* atau α.
Langkah 3	Hitung nilai uji dari data sampel dan tentukan nilai kritis yang diperoleh dari *significant level* yang telah ditentukan
Langkah 4	Bandingkan dua ukuran tahapan diatas. Apabila nilai uji dari data sampel berada di daerah kritis maka kesimpulannya adalah menolah dugaan awal dan sebaliknya jika nilai hasil uji pada data sampel berada di dalam daerah penerimaan maka kesimpulannya adalah gagal menolak dugaan awal.
Langkah 5	Menginterpretasikan dari kesimpulan yang didapat.

Pada sub bab berikut ini maka kita akan memulai melakukan uji inferensial yang paling sederhana yaitu uji z dan uji t hanya untuk satu sampel.

2. Uji Z dan uji T pada satu sampel

Baik uji Z dan uji T pada satu sampel bertujuan untuk menguji mean dari suatu populasi. Ketika σ dari populasi diketahui maka secara otomatis menggunakan pendekatan uji Z. Pada kenyataannya untuk mengetahui ukuran penyebaran dari suatu populasi sangatlah susah sehingga data penelitian banyak yang σ dari populasinya tidak diketahui terlebih jika banyaknya sampel juga kurang dari 30. Untuk mengatasi hal tersebut maka dapat digunakan uji T. Akan tetapi, walo σ dari populasi tidak diketahui akan tetapi jumlah sampel sangat banyak maka pendekatannya tetap menggunakan uji Z.

Pada sub bab ini yang akan kita bahas secara detail yang pertama adalah uji Z. Illustasinya adalah sebagai berikut :

Contoh 5.17. Ratna memberi tahu Riki bahwa skor rata-ratanya adalah 275. Riki skeptis dan meminta pembenaran atas perny-

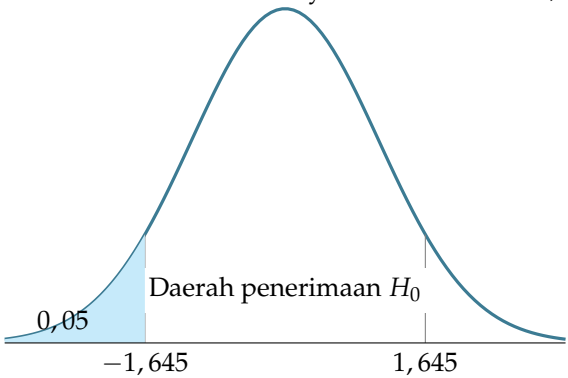
ataan dari Ratna. Oleh karena itu untuk membuktikan maka Ratna melakukan sebanyak 25 kali ujian. Rata-rata (sampel) dari 25 uji coba Jack hanya 264,4. Ratna masih berpendapat bahwa, ia kemampuan kinerjanya adalah 275 dan bahwa jika skornya yang relatif buruk itu karena hanya suatu kebetulan. Populasi standar deviasi adalah 20. Pada tingkat signifikansi 5%, apakah data memberikan bukti yang cukup untuk menyimpulkan bahwa skor rata-rata Jack kurang dari 275?. Gunakan $\alpha = 0,05$.

Data :
$266, 254, 248, 249, 297, 261, 293, 261, 266, 279, 222, 212, 282, 281, 265$
dan $240, 284, 253, 274, 243, 272, 279, 261, 273, 295$

Penyelesaian Gunakan langkah-langkah uji hipotesis dengan menggunakan *significant level* (α). ◀

Langkah 1 Buatlah hipotesisnya. $H_0 : \mu = 275$ dan $H_1 : \mu < 275$

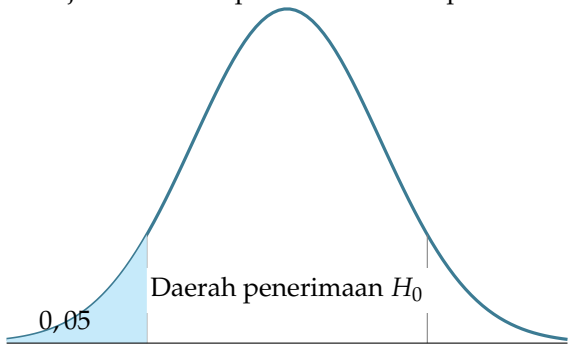
Langkah 2 Karena $\alpha = 0,05$ dari tabel Z diperoleh bahwa nilai kritisnya adalah $\pm 1,645$.



Langkah 3 Hitung nilai ujinya. $Z_{hitung} = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = -2,65$

Pendekatan diatas adalah menggunakan α sebagai alat untuk mengambil keputusan. Seperti yang telah dibahas diawal, maka contoh berikut akan mengilustrasikan bagaimana jika *P-value* digunakan pada soal yang sama.

Langkah 4 Buat keputusan. Tolak H_0 karena nilai uji berada pada daerah penolakan



H_0.

Langkah 5 Terdapat fakta yang kuat untuk menolak H_0. Sehingga dapat disimpulkan bahwa skor rata-rata Jack kurang dari 275

Contoh 5.18. Analisis data tentang skor Ratna dengan menggunakan pendekatan *P-value*.

Penyelesaian Pada intinya proses hampir sama dengan menggunakan α sebagai alat untuk menentukan keputusan. ◀

Langkah-langkah uji hipotesis dengan menggunakan *p-value*	
Langkah 1	Buatlah hipotesisnya. $H_0 : \mu = 275$ dan $H_1 : \mu < 275$
Langkah 2	Menghitung *P-value* dengan cara $$P(z < -2,65) = 0,004$$
Langkah 3	Tentukan nilai dari α yang digunakan, misalnaya $\alpha = 0,05$
Langkah 4	Bandingkan nilai hasil *P-value* dan α.
Langkah 5	Karena *P-value* lebih kecil dari α maka ada fakta yang kuat untuk menolak H_0. Dengan kata lain bahwa skor rata-rata Jack kurang dari 275

Berikut ini adalah R script untuk uji Z satu sampel dengan

menggunakan R studio

```
library(TeachingDemos)
datanew<-c(266, 254, 248,249,297,261, 293,261,266,279,
222,212,282,281,265,240,284,253,274,243,272,
279,261,273,295 )

### z test package
z.test(datanew,275,alternative="less",20).
```

Sedangkan hasil output dari R script diatas adalah sebagai berikut :

```
One Sample z-test

data:  datanew
z = -2.65, n = 25, Std. Dev. = 20, Std. Dev. of
the sample mean = 4, p-value = 0.004025
alternative hypothesis: true mean is less than 275
95 percent confidence interval:
-Inf 270.9794
sample estimates:
mean of datanew
264.4
```

Dari output menggunakan R studio diatas, dapat diketahui bahwa $Z_{hitung} = -2,65$ dan nilai dari $P - value = 0,004$. Kedua hasil tersebut sama persis dengan cara manual sebelumnya. Sehingga sama seperti dengan cara manual dapat disimpulkan bahwa ada fakta yang kuat untuk menolak H_0 yang artinya skor rata-rata Jack kurang dari 275.

Tidak setiap data dilapangan memiliki σ dari populasinya dan terutama apabila data sampel kurang dari 30, maka data yang demikian dapat dianalisis dengan uji T yang illustasikan akan dibahas berikut ini.

Contoh 5.19. Hujan asam dari pembakaran bahan bakar fosil telah menyebabkan banyak danau di seluruh dunia menjadi asam. Biologi di danau-danau ini sering runtuh karena perubahan kimia air yang cepat dan tidak menguntungkan. Danau diklasifikasikan sebagai non asam jika memiliki pH lebih besar dari 6. Gunakan

$\alpha = 0,05$.

Data :
7,2 7,3 7,3 6,3 5,7 6,9 6,1 5,5 6,7 6,9 6,3 7,9 dan 6,6 6,5 5,8

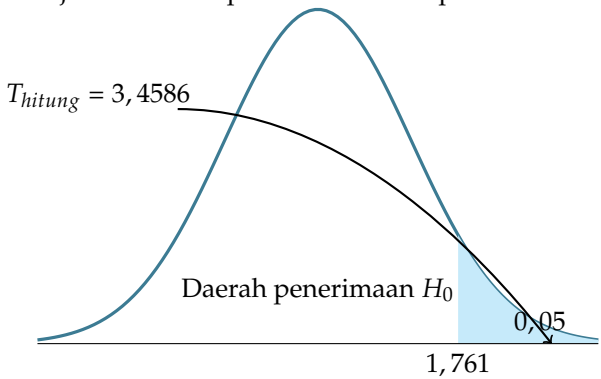
Penyelesaian Gunakan langkah-langkah uji hipotesis dengan menggunakan *significant level* (α). ◀

Langkah 1 Buatlah hipotesisnya. $H_0 : \mu = 6$ dan $H_1 : \mu > 6$

Langkah 2 Karena $\alpha = 0,05$ dan derajat kebebasannya ($df = n - 1 = 15 - 1 = 14$) dari tabel T diperoleh bahwa nilai kritisnya adalah $1,761$.

Langkah 3 Hitung nilai ujinya. $T_{hitung} = \dfrac{\bar{x} - \mu}{s/\sqrt{n}} = 3,4586$

Langkah 4 Buat keputusan. Tolak H_0 karena nilai uji berada pada daerah penolakan



H_0.

Langkah 5 Terdapat fakta yang kuat untuk menolak H_0. Danau dapat diklasifikasikan sebagai non asam karena memiliki pH lebih besar dari 6.

Sedangkan untuk menganalisis data tersebut menggunakan pendekatan *P-value* ada dicontoh berikutnya.

Contoh 5.20. Analisis data tentang danau dengan menggunakan pendekatan *P-value*.

Penyelesaian Pada intinya proses hampir sama dengan menggunakan α sebagai alat untuk menentukan keputusan. Langkah langkahnya terdapat pada Tabel 5.1 dibawah ini. ◀

Langkah-langkah uji hipotesis dengan menggunakan *p-value*	
Langkah 1	Buatlah hipotesisnya. $H_0 : \mu = 6$ dan $H_1 : \mu > 6$
Langkah 2	Menghitung *P-value* dengan cara $$T_{\alpha,df} = T_{\alpha,14} = 3,4586$$ dimana α diatas adalah peluangnya dari data sampel. Dari Tabel T didapat bahwa nilai peluang (*P-value*) lebih kecil $0,0025$
Langkah 3	Tentukan nilai dari α yang digunakan, misalnaya $\alpha = 0,05$
Langkah 4	Bandingkan nilai hasil *P-value* dan α.
Langkah 5	Karena *P-value* lebih kecil dari α maka ada fakta yang kuat untuk menolak H_0. Dengan kata lain bahwa danau dapat diklasifikasikan sebagai non asam karena memiliki pH lebih besar dari 6.

Tabel 5.1: Penyelesaian Uji T satu sampel dengan menggunakan pendekatan *P˙value*

Secara software, maka R script untuk uji T satu sampel soal diatas adalah sebagai berikut :

```
###### t test package
x<-c(7.2, 7.3, 7.3, 6.3, 5.7, 6.9, 6.1,
5.5, 6.7, 6.9, 6.3, 7.9, 6.6, 6.5, 5.8)
t.test(x, alternative="greater", mu=6).
```

Sedangkan hasilnya (output) dari R studio adalah

```
One Sample t-test

data:  x
t = 3.4586, df = 14, p-value = 0.001919
alternative hypothesis: true mean is greater than 6
95 percent confidence interval:
6.294448      Inf
```

```
sample estimates:
mean of x
6.6
```

Dari hasil diatas, didapat bahwa $T_{hitung} = 3,4586$ dan *p-value* adalah $0,00019$. Sehingga dapat dikatakan bahwa dengan cara manual dan software didapatkan hasil yang sama. Hasil pada software fokus pada nilai *P-value* yang nilainya kurang dari α sehingga ada fakta kuat untuk menolak H_0 atau bisa dikatakan significant. Dengan kata lain bahwa danau dapat diklasifikasikan sebagai non asam karena memiliki pH lebih besar dari 6.

3. Uji Z dan uji T pada dua sampel

Sub bab ini adah perluasan dari sub bab sebelumnya, yaitu dengan menambahkan menjadi dua sampel data.

Yang pertama kita fokus pada uji z dengan dua sampel. Illustrasinya adalah

Contoh 5.21. Data penelitian terhadap serum asam level pada Down's syndrome (groups 1) dan group normal (group 2) adalah sebagai berikut :

$$\bar{x}_1 = 9,78; n_1 = 900; \sigma_1 = 4,05$$

$$\bar{x}_2 = 15,10; n_2 = 1000; \sigma_2 = 4,28$$

Pertanyaan : apakah ada perbedaan antara rata rata dari individu yagn Down syndrom dengan normal?. Gunakan $\alpha = 0,05$.

Penyelesaian Gunakan langkah-langkah uji hipotesis adalah sebagai berikut : ◀

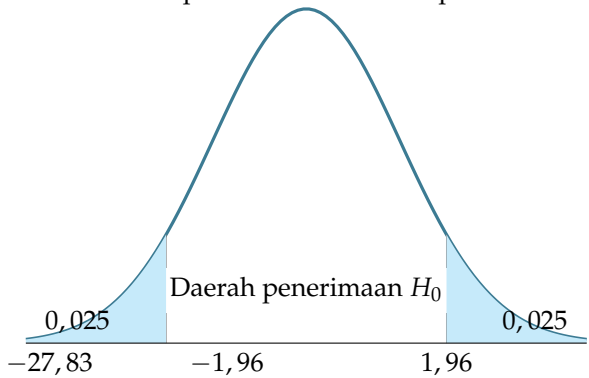
Langkah 1 Buatlah hipotesisnya. $H_0 : \mu_1 = \mu_2$ dan $H_1 : \mu_1 \neq \mu_2$

Langkah 2 Karena $\alpha = 0,05$ dari tabel Z diperoleh bahwa nilai kritisnya adalah $\pm 1,96$.

Langkah 3 Hitung nilai ujinya.

$$Z_{hitung} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = -27,83$$

Langkah 4 Buat keputusan.Tolak H_0 karena nilai uji berada pada daerah penolakan H_0.



Langkah 5 Terdapat fakta yang kuat untuk menolak H_0. Sehingga dapat disimpulkan bahwa ada perbedaan antara rata rata dari individu yang Down syndrom dengan normal

Contoh berikutnya adalah jika data tidak diketahui σ^2 atau varians dari dua populasi sehingga menggunakan uji T untuk dua sampel. Studi kasus yang pertama adalah jika dua populasi saling lepas (*independent*)

Contoh 5.22. Kepala departemen matematika tertarik pada perbedaan kinerja keterampilan menggambar tiga dimensi antara siswa tahun pertama yang diajar oleh guru yang berbeda. Para siswa yang masuk secara acak ditugaskan ke salah satu dari dua guru matematika dan diberikan tes kinerja standar setelah semester pertama. Kami mengambil sampel sembilan siswa dari satu kelas dan delapan dari lainnya.

Data kelas 1 adalah $35, 51, 66, 42, 37, 46, 60, 55, 53$
Data kelas 2 adalah $52, 87, 76, 62, 81, 71, 55, 67$

Pertanyaan : apakah ada perbedaan antara rata rata skor dari dua kelas tersebut?. Gunakan $\alpha = 0,05$.

Jika kita cermati, maka dua kelas tersebut memiliki siswa yang berbeda sehingga pendekatan yang dibutuhkan adalah uji T dengan dua sampel untuk kasus yang saling lepas.

Penyelesaian Gunakan langkah-langkah uji hipotesis adalah sebagai berikut. ◀

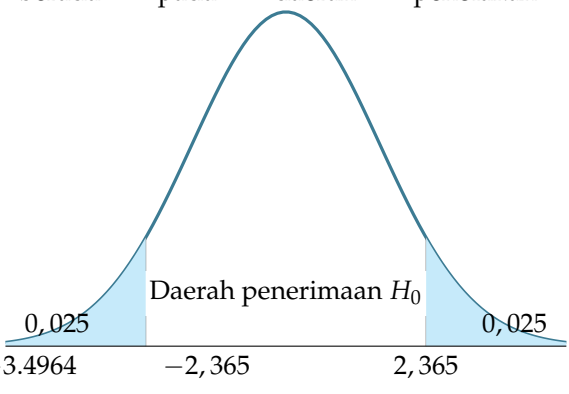
Langkah 1 Buatlah hipotesisnya. $H_0 : \mu_1 = \mu_2$ dan $H_1 : \mu_1 \neq \mu_2$

Langkah 2 Karena $\alpha = 0,05$ $df = n_{terendah} - 1 = 8 - 1 = 7$ dari tabel $T_{\frac{\alpha}{2},df} = T_{0,025,7}$ diperoleh bahwa nilai kritisnya adalah ± 2.365.

Langkah 3 Hitung nilai ujinya.

$$T_{hitung} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -3.4964$$

Langkah 4 Buat keputusan.Tolak H_0 karena nilai uji berada pada daerah penolakan H_0.



Langkah 5 Terdapat fakta yang kuat untuk menolak H_0. Sehingga dapat disimpulkan bahwa ada perbedaan antara rata rata skor dari dua kelas

Untuk analisis menggunakan R studio maka akan dijelaskan dengan R script berikut.

```
### T test untuk dua sampel sifatnya independent
data1 = c(35,51,66,42,37,46,60,55,53)
data2 = c(52,87,76,62,81,71,55,67)
t.test(data1,data2,alternative="two.sided", var.equal=FALSE)
```

Sedangkan hasil dari R studio adalah :

```
Welch Two Sample t-test

data:  data1 and data2
t = -3.4964, df = 13.823, p-value = 0.003623
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-31.364157  -7.496954
sample estimates:
mean of x mean of y
49.44444  68.87500.
```

Dari hasil output diatas, bisa kita dapatkan dua nilai yaitu $T_{hitung} = -3,4964$ dan *P-value*$=0,003623$. Untuk nilai T_{hitung} baik manual maupun software menghasilkan nilai yang sama. Sedangkan untuk software, kita fokus pada nilai *P-value* yang dihasilkan dan apat disimpulkan bahwa nilai *P-value*$=0,003623$ lebih dari α. Sehingga ada fakta untuk menolak H_0 yang artinya bahwa ada perbedaan antara rata rata skor dari dua kelas.

Berikut adalah illustrasi untuk kasus uji T dua sampel akan tetapi saling bergantung (*dependent*).

Contoh 5.23. Seorang guru kimia ingin menentukan keefektifan pelajaran praktikumnyanya. Dia memberikan tes keterampilan sederhana kepada sembilan siswa sebelum dimulainya kelas (pretest) dan tes keterampilan yang sama untuk siswa yang sama di akhir kelas (post test). Data adalah sebagai berikut :

Pretest adalah $78, 67, 56, 78, 96, 82, 84, 90, 87$
Post test adalah $80, 69, 70, 79, 96, 84, 88, 92, 92$

Pertanyaan: Apakah ada perbedaan prestasi dalam tes keterampilan antara pretest dan post test?. Gunakan $\alpha = 0,05$.

Penyelesaian Gunakan langkah-langkah uji hipotesis adalah sebagai berikut : ◀

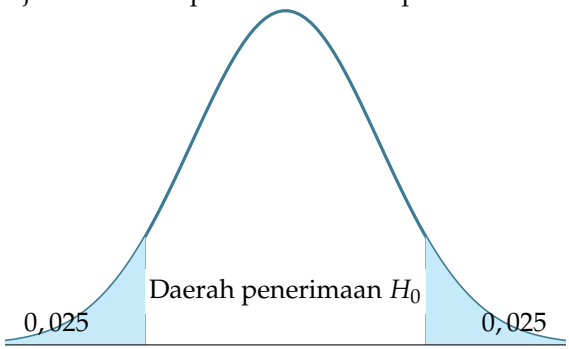
Langkah 1 Buatlah hipotesisnya. $H_0 : \mu_1 = \mu_2$ dan $H_1 : \mu_1 \neq \mu_2$ atau $H_0 : \delta = 0$ dan $H_1 : \delta \neq 0$

Langkah 2 δ = pretest-postest. Karena $\alpha = 0,05$ $df = n - 1 = 9 - 1 = 8$ dari tabel $T_{\frac{\alpha}{2},df} = T_{0,025,8}$ diperoleh bahwa nilai kritisnya adalah $\pm 2,306$.

Langkah 3 Hitung nilai ujinya.

$$T_{hitung} = \frac{\bar{\delta} - \delta}{\sqrt{\frac{\delta_s^2}{n}}} = -2,5478$$

Langkah 4 Buat keputusan. Tolak H_0 karena nilai uji berada pada daerah penolakan H_0.



Langkah 5 Terdapat fakta yang kuat untuk menolak H_0. Sehingga dapat disimpulkan bahwa ada perbedaan prestasi dalam tes keterampilan antara pretest dan post test.

Apabila kita ingin menganalisis menggunakan R studio, maka R scriptnya adalah sebagai berikut :

```
### T test untuk dua sampel yang sifatnya dependent
pre = c(78,67,56,78,96,82,84,90,87)
post = c(80,69,70,79,96,84,88,92,92)
```

```
t.test(pre,post,alternative="two.sided",paired=TRUE,mu=0)
```

dan hasil outputnya adalah

```
Paired t-test

data:  pre and post
t = -2.5478, df = 8, p-value = 0.03429
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-6.7736757 -0.3374354
sample estimates:
mean of the differences
-3.555556
```

Analisis data dengan uji T pada dua sampel yang sifatnya *dependent* bisa juga disebut sebagai uji T berpasangan (*paired T test*). Dari output tersebut nilai T_{hitung} baik cara manual maupun software menghasilkan nilai yang sama, yaitu $-2,5478$. Sedangkan *P-value*nya adalah $0,03429$ yang lebih kecil dari $\alpha = 0,05$, sehingga terdapat fakta yang kuat untuk menolak H_0. Hal tersebut dapat diartikan bahwa ada perbedaan prestasi dalam tes keterampilan antara pretest dan post test

4. Rangkuman

Rangkum untuk bab ini adalah yang pertama tentang prosedur uji hipotesis pada statistika inferensial yang terdapat pada Tabel 5.2.

Baik uji Z dan uji T pada satu sampel bertujuan untuk menguji mean dari suatu populasi.

Ketika σ dari populasi diketahui maka secara otomatis menggunakan pendekatan uji Z.

σ dari populasinya tidak diketahui terlebih jika banyaknya sampel juga kurang dari 30 maka dapat digunakan uji T.

Akan tetapi, walo σ dari populasi tidak diketahui akan tetapi jumlah sampel sangat banyak maka pendekatannya tetap menggunakan uji Z.

Langkah-langkah uji hipotesis pada statistika inferensial	
Langkah 1	Menentukan dugaan awal dan alternatifnya.
Langkah 2	Menentukan *significant level* atau α.
Langkah 3	Hitung nilai uji dari data sampel dan tentukan nilai kritis yang diperoleh dari *significant level* yang telah ditentukan
Langkah 4	Bandingkan dua ukuran tahapan diatas. Apabila nilai uji dari data sampel berada di daerah kritis maka kesimpulannya adalah menolah dugaan awal dan sebaliknya jika nilai hasil uji pada data sampel berada di dalam daerah penerimaan maka kesimpulannya adalah gagal menolak dugaan awal.
Langkah 5	Menginterpretasikan dari kesimpulan yang didapat.

Tabel 5.2: Prosedur uji hipotesis pada statistika inferensial

C. Perlatihan

Kerjakan soal-soal berikut ini untuk mempertajam pemahaman Saudara!

Suatu eksperimen tentang kadar gula pada 30 orang dimana 15 adalah perempuan dan sisanya adalah laki laki.

Data sebelum dilakukan treatment
Male : 134, 103, 116, 113, 124, 120, 128, 122, 123, 108, 134, 108, 111, 125, 134, Female : 105, 118, 97, 104, 99, 131, 108, 112, 97, 104, 103, 114, 96, 126, 109.

Data sesudah treatment
Male : 134, 106, 110, 115, 122, 126, 130, 118, 125, 110, 138, 111, 115, 125, 130, Female : 103, 120, 95, 105, 104, 134, 105, 110, 100, 107, 100, 112, 98, 127, 106.
Dengan significant level 5 persen, apakah ada perbedaan rata rata gula darah pada

a. laki laki dan perempuan sebelum treatment

b. Sebelum dan sesudah perlakukan pada laki laki

D. Buku Bacaan

Bluman, A. G. 2018. *Elementary Statistics: A Step By Step Approach (10th Edition)*. New York: McGraw-Hill Education.

Walpole, M. 2013. *Probability and Statistics for Engineers and Scientist: Global Edition (9th Edition)*. Mumbai: Pearson.

Weiss, N. A. 2017. *Introductory Statistics: Global Edition (10th Edition)*. Harlow: Pearson.

BAB VI
ANOVA

A. Pendahuluan

1. Deskripsi Singkat

Dalam bab sebelumnya kita telah mempelajari uji mean dari suatu populasi baik dari satu maupun dua sampel. Apabila terdapat lebih dari dua sampel maka uji Z maupun T tidak bisa lagi dapat digunakan maka, sehingga memerlukan pendekatan baru yaitu analisis varians (Anova).

2. Tujuan Pembelajaran

Menjelaskan pengetahuan tentang Anova satu arah

3. Anova satu arah

Seperti yang telah dijelaskan diatas bahwa anova adalah perluasan dari uji T ketika data yang diteliti oleh dari dua group. Pada bab ini akan difokuskan hanya pada analisis varians satu arah (*Anova one way*), dimaana hanya terdapat satu input (faktor).

Illustrasinya adalah sebagai berikut : Seorang peneliti ingin menentukan apakah merek sereal yang berbeda menyebabkan peningkatan jumlah berat yang berbeda pula. Subjek penelitian adalah tikus laboratorium. Para peneliti akan mulai dengan 20 bayi tikus dan empat merek sereal yang berbeda. Ada beberapa hal yang dilakukan yaitu : (1) secara acak membagi hewan menjadi empat kelompok, (2) mengkondisi lab yang identik, (3) yang berbeda hanya jenis merek sereal dalam setiap group.

Buatlah partisi jika varians dari semua pengamatan menjadi dua sumber variasi, yaitu variasi antara kelompok dan variasi dalam kelompok. Dari Tabel 6.1, dapat diketahui bahwa varians dalam kelompok (group) A adalah tidak ada (nol).

Kemudian group B telah dihitung kenaikan berat badannya dan data dapat dilihat di Tabel 6.2.

A	B	C	D
1			
1			
1			
1			

Tabel 6.1: Kenaikan berat badan (oz) tikus memberi makan empat merek sereal yang berbeda

A	B	C	D
1	7		
1	7		
1	7		
1	7		

Tabel 6.2: Kenaikan berat badan (oz) tikus memberi makan empat merek sereal yang berbeda

Dari Tabel 6.2, dapat diketahui bahwa varians dalam grup B adalah nol. Akan tetapi apabila kita cermati antara kelompok A dan kelompok B, maka terdapat perbedaan antar kelompok yang signifikan. Sehingga perbandingan antara varians dalam-grup (s_w^2) dan varians antara-grup (s_b^2) adalah dasar untuk prosedur Anova dan $\frac{s_b^2}{s_w^2}$ akan mengikuti distribusi F.

Proses tersebut dapat dirangkum sebagai berikut : dimana

	Sum of Squares	df	Mean Squares	F ratio	Critical F
Antar group	SS_b	$k-1$	$MS_b = \frac{SS_b}{k-1}$	$\frac{MS_b}{MS_w}$	$F_{(k-1),(N-k)}$
Dalam group	SS_w	$N-k$	$MS_w = \frac{SS_w}{N-k}$		
Total	SS_t	$N-1$			

Tabel 6.3: Anova satu arah

$$SS_b = \left(\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \ldots \right) - \left(\frac{(\sum_{i=1}^{k} \sum_{j=1}^{n} x)^2}{N} \right)$$

dan

$$SS_w = \sum_{i=1}^{k}\sum_{j=1}^{n} x^2 - \left(\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \ldots\right).$$

Contoh 6.24. Data lengkap dari kenaikan berat badan anak tikus ada di Tabel 6.4 sebagai berikut. Pertanyaannya : apakah merek

A	B	C	D
1	7	9	8
1	7	6	6
1	7	5	4
1	7	3	1
1	7	2	1

Tabel 6.4: Kenaikan berat badan (oz) tikus memberi makan empat merek sereal yang berbeda

sereal yang berbeda menyebabkan peningkatan jumlah berat yang berbeda pula?.

Penyelesaian Gunakan langkah-langkah uji hipotesis adalah sebagai berikut : ◀

Langkah 1 Buatlah hipotesisnya. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ dan H_1: paling tidak ada μ yang tidak sama dengan 0.

Langkah 2 Karena $\alpha = 0,05$, F_{df_b,df_w} dimana $df_b = k - 1$ dan $df_w = N - k$, k adalah banyaknya groups dan N adalah total pengamatan. Dari tabel $F_{df_b,df_w} = F_{3,16}$ diperoleh bahwa nilai kritisnya adalah .

Langkah 3 Hitung nilai ujinya.

$$F_{hitung} = \frac{MS_b}{MS_w} = 7,359$$

	Sum of Squares	df	Mean Squares	F ratio	Critical F
Antar group	93,75	3	31,25	7,3529	3,24
Dalam group	68,00	16	4,25		

Tabel 6.5: Anova satu arah

Langkah 4 Buat keputusan. Tolak H_0 karena nilai uji berada pada daerah penolakan H_0.

Langkah 5 Terdapat fakta yang kuat untuk menolak H_0. Sehingga dapat disimpulkan bahwa paling tidak ada μ yang tidak sama dengan 0. Dengan kata lain bahwa merek sereal yang berbeda menyebabkan peningkatan jumlah berat yang berbeda pula.

R script untuk menganalisis anova satu arah dengan R studio dengan menggunakan data diatas adalah sebagai berikut :

```
bb<-c(1,1,1,1,1,7,7,7,7,7,9,6,5,3,2,8,6,4,1,1)
group<-c("A","A","A","A","A","B","B","B","B","B",
"C","C","C","C","C", "D","D","D","D","D")
datanew<-data.frame(bb,group)
datanew
an<-lm(bb ~ group, data=datanew)
anova(an)
```

sedangkan output dari R studio adalah sebagai berikut :

```
Analysis of Variance Table

Response: bb
Df Sum Sq Mean Sq F value   Pr(>F)
group      3  93.75   31.25  7.3529 0.002576 **
Residuals 16  68.00    4.25
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1.
```

Dari hasil software, diketahui bahwa *p-value*nya adalah $0,0025$ yang artinya kurang dari $\alpha = 0,005$. Oeh karena itu dapat disimpulkan bahwa ada fakta kuat yang mendukung untuk menolak H_0. Dengan kata lain maka dapat diartikan bahwa merek sereal yang berbeda menyebabkan peningkatan jumlah berat yang berbeda pula. Dari uraian diatas baik menggunakan cara manual maupun software, menghasilkan hasil keputusan yang sama.

4. Rangkuman

Analisis varians merupakan perluasan dari uji T dan digunakan apabila terdapat lebih dari dua group dan anova satu arah berarti faktor (input) dari data hanya ada satu

Varians dalam grup (s_w^2) dan antar grup (s_b^2) menjadi hal yang penting dalam anova

$\frac{s_b^2}{s_w^2}$ akan mengikuti distribusi F_{df_b,df_w}.

B. Perlatihan

(1) Tabel 6.6 berikut menyajikan data berikut menyajikan data tentang umur dari tiga komunitas yang berbeda?.

komunitas A	komunitas B	komunitas C
16	65	45
15	43	30
25	77	22
30	90	66
39	82	47
20	69	33
16	73	50

Tabel 6.6: Umur dari tiga komunitas yang berbeda

(2) Tabel 6.7 berikut menyajikan data untuk pekerja yang dosis radiasi kumulatifnya dinilai di tiga lokasi berbeda. Tentukan apakah ada perbedaan yang signifikan dalam tingkat dosis rata-rata di antara tiga lokasi?.

lokasi A	lokasi B	lokasi C
11	29	37
27	41	51
19	19	42
21	39	28
31	24	35
14	35	48
28	46	75
22	64	49
18	52	61
10	23	52

Tabel 6.7: kumulatif radiokatif dari tiga lokasi yang berbeda

C. Buku Bacaan

Bluman, A. G. 2018. *Elementary Statistics: A Step By Step Approach (10th Edition)*. New York: McGraw-Hill Education.

Walpole, M. 2013. *Probability and Statistics for Engineers and Scientist: Global Edition (9th Edition)*. Mumbai: Pearson.

Weiss, N. A. 2017. *Introductory Statistics: Global Edition (10th Edition)*. Harlow: Pearson.

BAB VII
KORELASI

A. Pendahuluan

1. Deskripsi Singkat

Pada bab sebelumnya, kita telah mempelajari bagaimana cara untuk pengujian hipotesis. Dalam bidang statistika yang lain, khususnya statistika inferensia juga terdapat suatu metode untuk menentukan apakah ada hubungan antara dua atau lebih variabel numerik atau kuantitatif.

2. Tujuan Pembelajaran

a. Menjelaskan pengetahuan tentang korelasi secara umum.
b. Menjelaskan pengetahuan tentang diagram pencar.
c. Menjelaskan pengetahuan tentang menghitung koefisien korelasi.
d. Menjelaskan pengetahuan tentang koefisien korelasi pada populasi dan uji hipotesisnya.
e. Menerapkan pengetahuan korelasi untuk data dependen dalam kehidupan sehari-hari.

B. Materi

1. Persamaan Korelasi

Korelasi adalah metode statistik yang digunakan untuk menentukan ada atau tidaknya hubungan linear antara variabel. Seringkali seorang pengusaha mungkin ingin tahu apakah volume penjualan untuk bulan tertentu terkait dengan jumlah iklan yang dilakukan perusahaan pada bulan itu. Seorang guru juga tertarik untuk menentukan apakah total jam belajar siswa berkaitan dengan skor siswa pada ujian mata pelajaran tertentu. Seorang peneliti di bidang medis tertarik pada pertanyaan seperti: apakah kafein terkait dengan kerusakan jantung? atau apakah ada hubungan antara usia seseorang dan tekanan darahnya? Seorang

ahli zoologi mungkin juga ingin tahu apakah berat lahir hewan tertentu terkait dengan rentang hidupnya.

Para statistikawan menggunakan ukuran numerik untuk menentukan apakah dua atau lebih variabel terkait secara linier dan untuk menentukan kekuatan hubungan antara atau di antara variabel. Ukuran ini disebut *Koefisien korelasi*. Misalnya, ada banyak variabel yang dapat menyebabkan penyakit jantung, di antaranya kekurangan olahraga, merokok, faktor keturunan, usia, tingkat stres, dan pola makan. Dari variabel-variabel ini, beberapa variabel lebih penting daripada yang lain.

Terdapat dua jenis hubungan, yaitu **sederhana** dan **berganda**. Dalam hubungan sederhana, terdapat dua variabel yaitu variabel independen yang dapat disebut sebagai variabel bebas dan variabel dependen yang sering disebut sebagai variabel terikat. Sebagai contoh, seorang manajer mungkin ingin melihat apakah tenaga penjualan yaang berpengalaman itu berhubungan dengan jumlah penjualan. Hubungan tersebut merupakan hubungan sederhana karena hanya ada dua variabel — pengalaman dan jumlah penjualan.

Dalam hubungan berganda, disebut juga regresi berganda, dua atau lebih variabel bebas digunakan untuk memprediksi suatu variabel terikat. Misalnya, seorang dosen ingin menyelidiki apakan terdapat hubungan antara kesuksesan di perguruan tinggi dengan beberapa faktor seperti lamanya belajar, nilai indeks prestasi, dan latar belakang pendidikan sebelumnya.

Hubungan sederhana dikategorikan menjadi dua: positif atau negatif. **Hubungan positif** terjadi ketika kedua variabelnya meningkat atau menurun dalam waktu yang sama, misalnya tinggi badan dan berat badan seorang anak. Tentu saja hal ini adalah hal yang saling berkaitan karena semakin tinggi seseorang akan semakin bertambah berat badannya. Hubungan sederhana seperti ini adalah hubungan yang positif. Sebaliknya, **hubungan negatif** terjadi jika salah satu variabel bertambah dan variabel lainnya berkurang, begitu juga sebaliknya. Hubungan ini paling mudah dijumpai pada kekuatan seseorang. Jika kita mengukur kekuatan seseoang yang berusia tujuh puluh tahun, biasanya kita akan mengetahui bahwa kekuatannya juga akan menurun.

2. Diagram Pencar dan Korelasi

Dalam korelasi sederhana, seorang ilmuwan mengumpulkan data dari dua variabel kuantitatif untuk melihat apakah ada hubungan antarkedua variabel tersebut. Misalnya pada tabel berikut ini: Variabel bebas adalah variabel yang dapat dikontrol

Tabel 7.1: Lama belajar dan Nilai ujian

Siswa	Lama Belajar	Nilai ujian
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

sedangkan variabel terikat adalah variabel yang tidak dapat dikontrol. Dalam kasus ini, variabel bebasnya adalah lama belajar dan dinyatakan sebagai variabel x. Variabel terikatya adalah nilai ujian dan dinyatakan sebagai variabel y. Alasan perbedaan antar-variabel adalah bahwa kita mengasumsikan bahwa nilai nilai yang diperoleh siswa tergantung pada jumlah jam yang dipelajari siswa. Kita juga mengasumsikan bahwa, sampai batas tertentu, siswa dapat mengatur atau mengontrol lama belajar untuk ujian.

Penentuan variabel x dan y tidak selalu jelas, dan seringkali ditetukan secara sebarang. Misalkan, jika seorang peneliti menyelidiki efek usia pada tekanan darah. Dia dapat saja mengasumsikan bahwa usia mempengaruhi tekanan darah sehingga variabel usia disebut variabel bebas dan variabel tekanan darah disebut variabel terikat. Jika seorang peneliti menyelidiki perilaku suami pada hal tertentu dan perilaku istri pada hal yang sama, akan sulit untuk menyatakan mana variabel bebas dan terikatnya. Dalam kasus tersebut, peneliti dapat mendesain mana yang merupakan variabel bebas dan variabel terikatnya secara bebas.

Variabel bebas dan terikat tersebut dapat digambar dalam suatu diagram yang disebut diagram pencar. Variabel bebas x digambar pada sumbu mendatar (sumbu-x) sedangkan variabel terikat y digambar pada sumbu tegak (sumbu-y).

Diagram pencar adalah suatu grafik dari pasangan terurut (x, y) dari bilangan yang terdiri dari variabel bebas x dan variabel terikat y.

Diagram pencar adalah suatu cara penyajian data secara visual mengenai hubungan antarvariabel bebas dan terikat. Skala-skala dalam variabelnya dapat saja berbeda dan koordinat sumbu ditentukan oleh nilai data terkecil dan terbesar dari variabel-variabelnya.

Langkah-langkah membuat diagram pencar	
Langkah 1	Gambarlah dan berilah label sumbu-x dan sumbu-y.
Langkah 2	Plot setiap data pada diagram dengan menandai diagram dengan suatu titik.

Contoh 7.25. Buatlah diagram pencar dari Tabel 7.1.

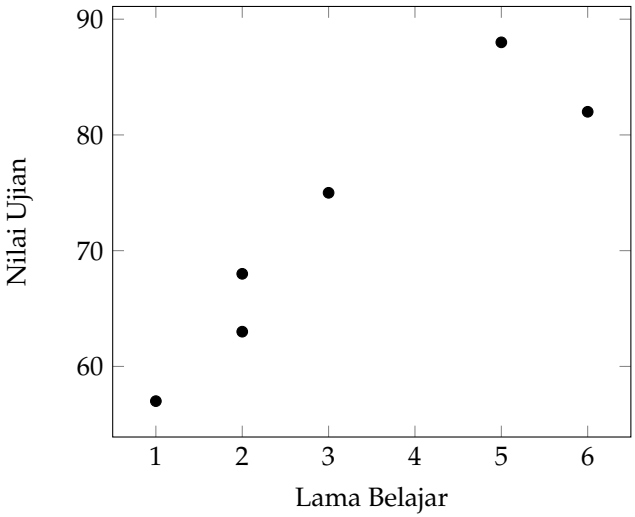
Penyelesaian Dari Tabel 7.1 diperoleh suatu diagram pencar sebagai berikut:

Source Code 7.1: *Syntax* R untuk membuat diagram pencar

```
x<-c(6,2,1,5,2,3) ## nilai dari variabel bebas
y<-c(82,63,57,88,68,75) ## nilai dari variabel terikat
plot(x,y) ## menggambar diagram pencar
```

◀

Setelah diagram tergambar, diagram tersebut harus dianalisis untuk menentukan bentuk hubungannya (jika ada). Misalnya pada contoh tersebut menunjukkan bahwa terjadi hubungan yang positif antara lama belajar dan nilai ujian. Semakin lama waktu belajar, semakin baik pula nilai ujian yang diperoleh. Perhatikan pula bahwa apabila titik-titik pada Gambar 7.1 dihubungkan, kurva yang terjadi berupa sebuah garis lurus meskipun tidak sempurna. Hal yang demikian ini kita sebut sebagai hubungan linear.



Gambar 7.1: Diagram pencar lama belajar vs. nilai ujian

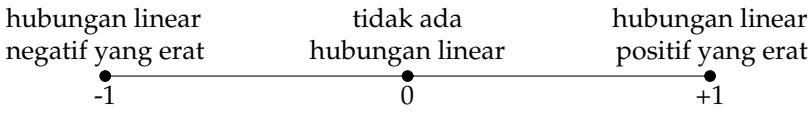
3. Koefisien Korelasi

Ada beberapa koefisien korelasi di dalam Statistika. Koefisien yang dibahas dalam bagian ini adalah Koefisien Korelasi *Product-Moment* Pearson (*Pearson Product-Moment Correlation Coefficient/PPMCC*) yang dinamakan dari penemunya, yaitu Karl Pearson.

> **Koefisien korelasi** yang dihitung dari ukuran sampel data mengukur keeratan dan arah hubungan linear dari dua variabel kuantitaif. Simbol dari koefisien korelasi sampel adalah r dan simbol dari koefisien korelasi populasi adalah ρ.

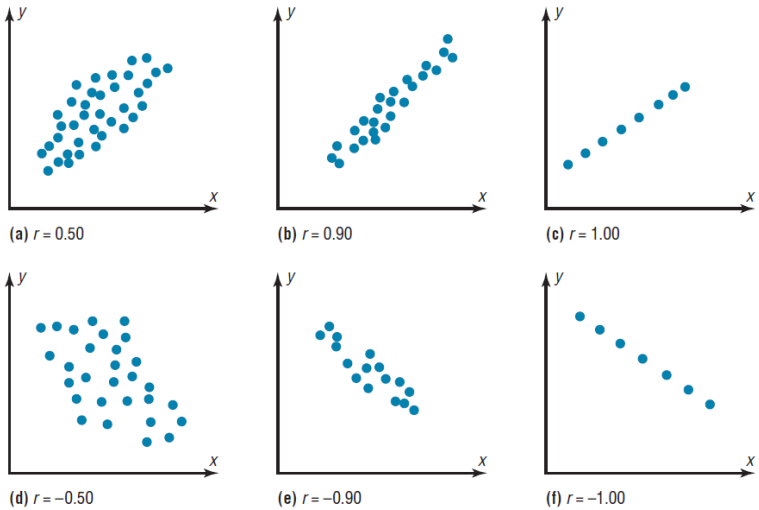
Nilai dari koefisien korelasi berkisar antara -1 hingga +1. Jika terdapat hubungan positif linear yang erat, nilai r akan dekat dengan +1. Sebaliknya, jika terdapat hubungan negatif linear yang erat, nilai r akan dekat dengan -1. Ketika tidak terdapat hubungan antarvariabel atau hanya hubungan yang lemah, nilai r akan mendekati 0 (perhatikan **??**).

Gambar 7.3 menunjukan hubungan antara koefisien korelasi



Gambar 7.2: Rentang nilai koefisien korelasi

dan diagram pencar. Ingat bahwa ketika nilai koefisien korelasinya naik dari 0 ke +1 (gambar a, b, dan c), nilai-nilainya menjadi dekat dengan suatu hubungan erat yang meningkat. Ketika nilai koefisiennya turun dari 0 ke -1 (gambar d, e, dan f), nilai-nilainya mejadi dekat dengan suatu garis lurus.



Gambar 7.3: Hubungan antara koefisien korelasi dan diagram pencar

Ketika menghitung koefisien korelasi, kita memperhatikan beberapa asumsi, yaitu

- Sampelnya acak.
- Pasangan data memenuhi garis lurus dan diukur pada level interval atau rasio.

Sifat-sifat koefisien korelasi linear	
(i)	Koefisien korelasi tidak memiliki satuan.
(ii)	Nilai r akan selalu berada diantara -1 dan +1, sehingga $\lvert r \rvert \leq 1$.
(iii)	Jika nilai x dan y saling ditukar satu dengan lain, nilai r tidak akan berubah.
(iv)	Jika nilai x dan/atau y dikonversi ke dalam skala yang lain, nilai r tidak akan berubah.
(v)	Nilai r sensitif terhadap pencilan dan dapat berubah secara drastis jika terdapat pencilan di suatu data.

- Variabel-variabelnya berdistribusi normal gabungan.

Ada beberapa rumus untuk menghitung koefisien korelasi, salah satu rumusnya adalah sebagai berikut:

$$\boxed{r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\left(\sum x^2\right) - \left(\sum x\right)^2\right]\left[n\left(\sum y^2\right) - \left(\sum y\right)^2\right]}}} \tag{7.1}$$

dimana n adalah jumlah data yang berpasangan.

Langkah-langkah menghitung koefisien korelasi	
Langkah 1	Buatlah tabel yang lengkap seperti contoh berikut:

x	y	xy	x^2	y^2

Langkah 2	Masukkan data ke kolom-kolom variabel yang bersesuaian.
Langkah 3	Jumlahkan total untuk setiap kolom, kemudian substituskan ke dalam rumus 7.1

Contoh 7.26. Carilah koefisien korelasi dari data pada Tabel 7.1

Penyelesaian Buatlah Tabel 7.1 menjadi tabel yang lengkap sebagai berikut:

Siswa	Lama Belajar (x)	Nilai ujian (y)	xy	x^2	y^2
A	6	82	492	36	6.724
B	2	63	126	4	3.969
C	1	57	57	1	3.249
D	5	88	440	25	7.744
E	2	68	136	4	4.624
F	3	75	225	9	5.625
$\sum$	19	433	1.476	79	31.935

Substitusilah nilai-nilai dari tabel tersebut ke dalam rumus 7.1

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\left(\sum x^2\right) - \left(\sum x\right)^2\right]\left[n\left(\sum y^2\right) - \left(\sum y\right)^2\right]}}$$

$$r = \frac{6\,(1.476) - (19)\,(433)}{\sqrt{\left[6\,(79) - (19)^2\right]\left[6\,(31.935) - (433)^2\right]}}$$

$$r = 0.9217$$

Koefisien korelasi menunjukkan adanya hubungan yang erat antara lama belajar dan nilai ujian siswa.

Source Code 7.2: *Syntax* R untuk menghitung koefisien korelasi

```
x<-c(6,2,1,5,2,3) ## nilai dari variabel bebas
y<-c(82,63,57,88,68,75) ## nilai dari variabel terikat
r<-cor(x,y) ## menghitung koefisien korelasi
```

◀

4. Pengujian Hipotesis Koefisien Korelasi

Ketika nilai r dekat dengan 0, hubungan linearnya lemah atau tidak ada. Karena nilai r dihitung dari data yang diperoleh dari sampel, dua kemungkinan ketika r yang terjadi tidak sama dengan nol adalah nilai r terlalu tinggi untuk dapat disimpulkan bahwa ada hubungan linear yang signifikan antara variabel-variabelnya atau nilai r adalah nilai yang kebetulan. Untuk mengambil keputusan terhadap nilai r, kita gunakan prosedur uji hipotesis.

Koefisien korelasi populasi dihitung dengan mengambil seluruh nilai pasangan (x, y) yang mungkin, notasinya dilam-

Langkah-langkah uji hipotesis	
Langkah 1	Buatlah hipotesis.
Langkah 2	Carilah nilai kritisnya.
Langkah 3	Hitunglah nilai ujinya.
Langkah 4	Ambillah keputusan berdasarkan hasil nilai uji dan nilai kritisnya.
Langkah 5	Tariklah kesimpulan.

bangkan dengan ρ. Koefisien korelasi sampel dapat digunakan sebagai estimator dari r jika asumsi-asumsi berikut valid:

- Variabel x dan y berhubungan linear.
- Variabel-variabelnya merupakan variabel acak.
- Dua variabelnya berdistribusi bivariat normal.

Distribusi bivariat normal adalah untuk pasangan nilai data (x, y), nilai-nilai y yang bersesuaian berdistribusi lonceng untuk sebarang x dan nilai x untuk sebarang y juga berdistribusi lonceng.

> Secara formal, **koefisien korelasi** ρ adalah korelasi yang dihitung dengan menggunakan nilai pasangan data (x, y) yang diambil dari populasi.

Dalam uji hipotesis, salah satu dari dua hal berikut ini adalah benar:

$\mathbf{H_0} : \rho = 0$ (tidak ada korelasi antara variabel x dan y)
$\mathbf{H_1} : \rho \neq 0$ (ada korelasi yang signifikan antara variabel x dan y)

Ketika hipotesis nol ditolak pada selang kepercayaan tertentu, hal ini bermakna bahwa ada perbedaan yang signifikan antara nilai r dan 0. Ketika hipotesis nol tidak ditolak, hal ini berarti bahwa nilai r tidak berbeda secara signifikan dari nol dan mungkin berupa kebetulan saja. Beberapa metode dapat digunakan untuk menguji signifikansi koefisien korelasi, salah satunya adalah dengan menggunakan uji-t.

$$\boxed{t = r\sqrt{\frac{n-2}{1-r^2}}} \tag{7.2}$$

dengan derajat kebebasan sama dengan $n - 2$.

Meskipun uji hipotesis dapat saja searah, kebanyakan hipotesis melibatkan koefisien korelasi yang dua arah. Ingat bahwa r menyatakan koefisien korelasi populasi. Jika tidak ada hubungan linear, nilai dari koefisien korelasi akan sama dengan nol sehingga hipotesis menjadi $H_0 : \rho = 0$ dan $H_1 : \rho \neq 0$.

Contoh 7.27. Ujilah signifikansi koefisien korelasi dari contoh sebelumnya. Gunakan $\alpha = 0,05$.

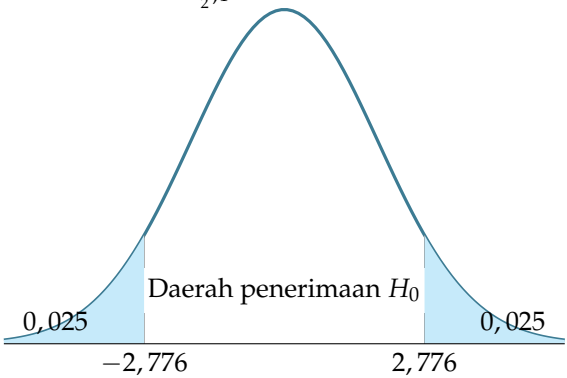
Penyelesaian Gunakan langkah-langkah uji hipotesis. ◀

Langkah 1 `Buatlah hipotesisnya.`

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Langkah 2 `Carilah nilai kritisnya dari tabel` t

Karena $\alpha = 0,05$ dan DK=4 maka diperoleh nilai kritis dari tabel $t_{\frac{\alpha}{2},4} = \pm 2,776$.

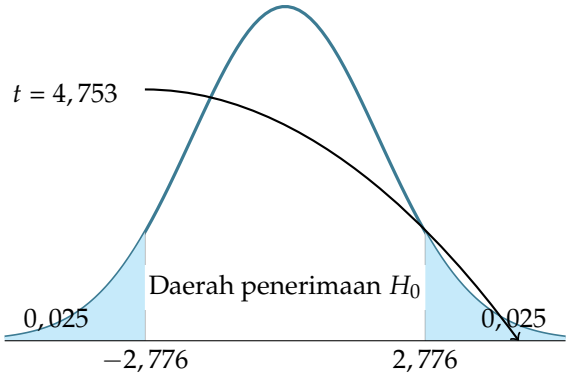


Langkah 3 `Hitung statistik ujinya.`

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0,9217\sqrt{\frac{4}{0,1503}} = 4,753$$

Langkah 4 `Buat keputusan berdasarkan statistik uji dan nilai kritis.`

Tolak H_0 karena nilai statistik uji berada pada daerah penolakan H_0.



Langkah 5 `Tariklah kesimpulan.`
Pada level signifikansi 5%, data tersebut memberikan cukup bukti untuk menyimpulkan bahwa koefisien garis singgungnya tidak sama dengan nol sehingga terdapat hubungan yang signifikan antara lama belajar dan hasil ujian siswa.

5. Rangkuman

Korelasi adalah bertujuan untuk mendapatkan hubungan antara dua variabel (x, y) dan nilainya terletak antara 0 sampai dengan 1. Korelasi juga bisa bernilai positif atau negatif.

Jika nilai korelasi adalah -1 maka artinya adalah hubungan tersebut sangat kuat namun saling berlawanan. Dengan kata lain bahwa x meningkat tetapi y menurun. Sebaliknya

jika korelasi bernilai 1 maka artinya hubungan tersebut sangat kuat dengan kondisi x naik maka y juga akan naik.

Apabila dua variabel telah memiliki korelasi maka langkah selanjutnya adalah memprediksi nilai y dari sebarang nilai x, yaitu dengan suatu teknis namanya adalah regresi yang akan dibahas pada bab selanjutnya.

C. Perlatihan

Kerjakan soal-soal berikut ini untuk mempertajam pemahaman Saudara!

(1) Seorang peneliti ingin melihat adanya hubungan antara jumlah kasus cacar dan parotitis yang dilaporkan selama lima tahun terakhir ini di Amerika Serikat. Berikut ini adalah data yang berhasil dihimpun oleh peneliti tersebut

Kasus cacar	43	140	71	63	112
Kasus parotitis	800	454	1991	2612	370

Adakah hubungan diantara keduanya?

(2) Suatu sampel acak mengenai harapan hidup pria dan wanita dari negara dunia ketiga dipilih. Berikut ini adalah datanya:

Pria	59,7	72,9	41,9	46,2	50,3	43,2
Wanita	63,8	77,8	44,5	48,3	54,0	43,5

Adakah hubungan diantara antara harapan hidup pria dan wanita di negara dunia ketiga?

(3) Seorang arsitek ingin menentukan hubungan antara ketinggan bangunan (dalam meter) dan banyak lantai di dalam bangunan. Dia telah menghimpun sampel dari delapan gedung di Kota Chicago seperti pada tabel berikut ini:

Jumlah lantai	64	68	50	48	32	46	58	45
Keting-gian	303	257	223	206	197	193	185	183

Jelaskan hubungan di antara keduanya!

D. Bahan Bacaan

Bluman, A. G. 2018. *Elementary Statistics: A Step By Step Approach (10th Edition)*. New York City: McGraw-Hill Education.

Walpole, M. 2013. *Probability and Statistics for Engineers and Scientist:*

Global Edition (9th Edition). Mumbai: Pearson.

Weiss, N. A. 2017. *Introductory Statistics: Global Edition (10th Edition)*. Harlow: Pearson.

BAB VIII
REGRESI

A. Pendahuluan

1. Deskripsi Singkat

Pada bab sebelumnya kita telah membahas bagaimana cara menghitung koefisien korelasi, menentukan hubungan linear di antara dua variabel, hingga menguji signifikansi hubungannya. Jika nilai koefisien korelasinya signifikan, maka langkah selanjutnya adalah menentukan **garis regresi**nya, yang merupakan garis yang paling sesuai. Apabila r tidak signifikan, kita tidak perlu menentukan garis regresinya. Tujuan menentukan garis regresi adalah untuk memungkinkan kita melihat suatu tren dan membuat prediksi berdasarkan data.

2. Tujuan Pembelajaran

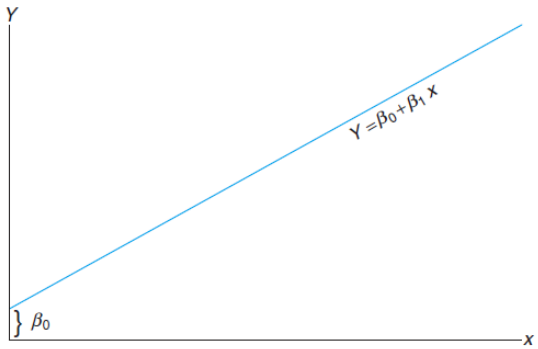
a. Menjelaskan pengetahuan tentang penentuan variabel bebas dan variabel terikat.
b. Menjelaskan metode estimasi parameter dalam regresi.
c. Menjelaskan pengetahuan tentang uji hipotesis untuk koefisien regresi.
d. Menjelaskan pengetahuan tentang uji hipotesis untuk konstanta regresi.
e. Menerapkan pengetahuan regresi linear dalam kehidupan sehari-hari.

B. Materi

1. Persamaan Regresi

Regresi adalah metode statistik yang digunakan untuk menggambarkan sifat hubungan antara variabel, yaitu positif atau negatif, linear atau nonlinear. Suatu hubungan yang masuk akal dari respons y dan regresor x adalah hubungan linear.

$$y = \beta_0 + \beta_1 x$$



Gambar 8.1: Suatu hubungan linear

dimana β_0 adalah pemotong dan β_1 adalah kemiringan. Konsep analisis regresi berkaitan dengan menemukan hubungan terbaik antara variabel y dan x, mencari mengukur kekuatan hubungan itu, dan menggunakan metode yang memungkinkan prediksi nilai variabel terikat yang diberikan nilai-nilai regressor x. Dalam banyak aplikasi, akan ada lebih dari satu regresi (lebih dari satu variabel bebas yang membantu menjelaskan y). Misalnya, dalam kasus di mana variabel terikatnya adalah harga sebuah rumah, orang akan berpikir bahwa usia rumah dan luas bangunan berpengaruh pada harga, sehingga dalam hal ini struktur regresi berganda dapat ditulis sebagai

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

di mana y adalah harga, x_1 adalah luas tanah, dan x_2 adalah usia rumah dalam tahun. Persamaan yang dihasilkan disebut regresi berganda, sedangkan analisis kasus regresi tunggal disebut regresi sederhana. Sebagai ilustrasi kedua dari regresi berganda adalah sebagai berikut: Seorang insinyur kimia mungkin memperhatikan jumlah hidrogen yang hilang dari sampel logam tertentu ketika bahan tersebut disimpan. Dalam hal ini, mungkin ada dua input, waktu penyimpanan x_1 dalam jam dan suhu penyimpanan x_2 dalam derajat celcius. Variabel terikatnya adalah hidrogen yang hilang y dalam bagian per juta.

2. Regresi Linear Sederhana

Sebuah analisis hubungan antara y dan x membutuhkan model statistik. Suatu model sering digunakan oleh statistikawan sebagai representasi suatu kondisi ideal yang pada dasarnya mendefinisikan bagaimana kita memahami bahwa suatu data yang dihasilkan oleh sistem yang bersangkutan. Model statistik tersebut harus mencantumkan himpunan $\{(x_i, y_i); i = 1, 2, \cdots, n\}$ data yang melibatkan n pasangan nilai (x, y). Kita harus ingat bahwa nilai y_i bergantung pada x_i melalui struktur linear yang juga melibatkan komponen acak. Dasar penggunaan model statistik berkaitan dengan bagaimana variabel acak y bergerak dengan x dan komponen acak. Model ini juga mencakup apa yang diasumsikan tentang sifat statistik dari komponen acak. Model statistik untuk regresi linier sederhana diberikan oleh:

$$\boxed{y = \beta_0 + \beta_1 x + \epsilon} \tag{8.1}$$

Pada persamaan 8.1, β_0 dan β_1 adalah parameter perpotongan dan kemiringan. ϵ adalah suatu variabel acak yang diasumsikan berdistribusi dengan $\mathbf{E}(\epsilon) = 0$ dan $\text{Var}(\epsilon) = \sigma^2$. Nilai dari σ^2 sering disebut sebagai varians galat atau varians residual. Distribusi normal dari ϵ, $n\left(x; 0, \sigma^2\right)$.

> **Garis regresi** adalah suatu garis yang paling sesuai dengan sekumpulan data berdasarka kriteria kuadrat terkecil.
> **Persamaan regresi** adalah persamaan dari garis regresi.

Dari model tesebut, beberapa hal menjadi jelas. Variabel y adalah variabel acak karena ϵ acak. Nilai x dari variabel regresor tidak acak dan pada kenyataannya, diukur dengan galat yang dapat diabaikan. Nilai ϵ sering disebut sebagai **galat acak** atau **gangguan acak**, memiliki varians yang konstan. Asumsi ini sering disebut **asumsi varians homogen**.

Aspek penting dalan analisis regresi sangat sederhana, yaitu mengestimasi parameter β_0 dan β_1. Misalkan kita mengestimasi β_0 dengan $\hat{\beta}_0$ dan β_1 dengan $\hat{\beta}_1$, estimasi dari garis regresinya menjadi

$$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x} \tag{8.2}$$

dimana $\hat{y}$ adalah nilai yang diprediksi. Jelas bahwa garis yang sesuai adalah estimasi dari garis regresi yang sebenarnya. Nilai-nilai estimator b_0 dan b_1 dicari dengan menggunakan rumusan sebagai berikut:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} \tag{8.3}$$

$$\hat{\beta}_0 = \frac{1}{n}\left[\sum_{i=0}^{n} y_i - \hat{\beta}_1 \sum_{i=0}^{n} x_i\right] \tag{8.4}$$

dengan:

Kuantitas	Definisi Rumus	Cara Menghitung
s_{xx}	$\sum_{i=0}^{n} (x_i - \bar{x})^2$	$\sum_{i=0}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=0}^{n} x_i\right)^2$
s_{xy}	$\sum_{i=0}^{n} (x_i - \bar{x})(y_i - \bar{y})$	$\sum_{i=0}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=0}^{n} x_i\right)\left(\sum_{i=0}^{n} y_i\right)$
s_{yy}	$\sum_{i=0}^{n} (y_i - \bar{y})^2$	$\sum_{i=0}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=0}^{n} y_i\right)^2$

Langkah-langkah menghitung persamaan regresi

Langkah 1 Buatlah tabel yang lengkap seperti contoh berikut:

x	y	xy	x^2	y^2

Langkah 2 Masukkan data ke kolom-kolom variabel yang bersesuaian.

Langkah 3 Jumlahkan total untuk setiap kolom, kemudian substituskan ke dalam rumus (8.4) dan (8.3)

Langkah 4 Subsitusikan nilai $\hat{\beta}_0$ dan $\hat{\beta}_1$ yang diperoleh dalam langkah sebelumnya ke persamaan (8.2)

Contoh 8.28. Carilah persamaan garis regresi untuk data penjualan mobil pada Tabel 8.1 dan gambarkan garis regresinya.

Tabel 8.1: Data penjualan mobil di Amerika Serikat pada tahun 2018

Perusahaan	Banyak mobil terjual (dalam puluhan ribu)	Pendapatan (dalam juta dolar)
A	63,0	7,0
B	29,0	3,9
C	20,8	2,1
D	19,1	2,8
E	13,4	1,4
F	8,5	1,5

Penyelesaian Dengan memisalkan banyak mobil yang terjual sebagai x dan pendapatan sebagai y, dan dengan melengkapi data pada Tabel 8.1, maka:

Tabel 8.2: Tabel lengkap data penjualan mobil

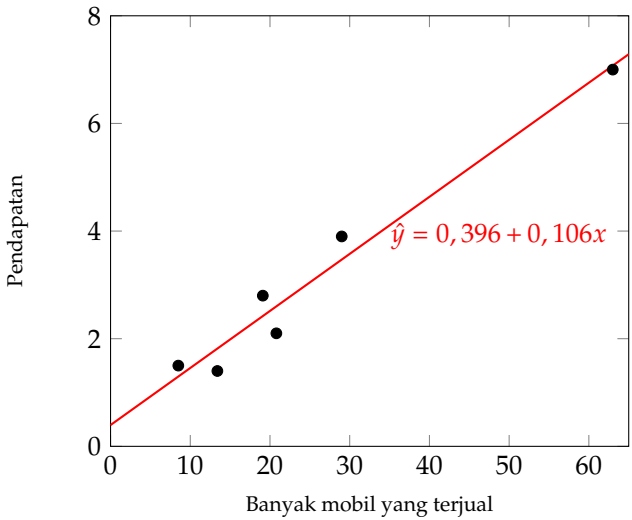
Perusahaan	x	y	xy	x^2	y^2
F	8,5	1,5	12,75	72,25	2,25
E	13,4	1,4	18,76	179,56	1,96
D	19,1	2,8	53,48	364,81	7,84
C	20,8	2,1	43,68	432,64	4,41
B	29	3,9	113,1	841	15,21
A	63	7	441	3969	49
$\sum$	153,8	18,7	682,77	5.859,26	80,67

dari Tabel 8.2 kita peroleh:

$$
\begin{aligned}
s_{xx} &= 1.916,853 \\
s_{xy} &= 203,427 \\
s_{yy} &= 22,388 \\
\hat{\beta}_1 &= \frac{s_{xy}}{s_{xx}} = \frac{203,427}{1.196,853} = 0,106 \\
\hat{\beta}_0 &= \frac{1}{6}\left[18,7 - 0,106\,(153,8)\right] = 0,396
\end{aligned}
$$

Persamaan garis regresinya adalah $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ adalah $\hat{y} = 0,396 + 0,106x$. Untuk membuat garis regresinya, kita ambil dua titik

diantara 10 dan 60. Misalkan $x = 15$, dengan menyubtitusikan $x = 15$ ke persamaan, kita peroleh $\hat{y} = 1,986$. Misalkan juga $x = 40$, kita peroleh $\hat{y} = 4,636$. Jika titik-titik $(15, 1,986)$ dan $(40, 4,636)$ di plot, maka akan diperoleh garis regresi sebagai berikut:



Gambar 8.2: Diagram pencar dan garis regresi

Source Code 8.1: *Syntax* R untuk membuat garis regresi

```
x<-c(63,29,20.8,19.1,13.4,8.5) ## nilai dari variabel bebas
y<-c(7,3.9,2.1,2.8,1.4,1.5) ## nilai dari variabel terikat
plot(x,y, ylab="Pendapatan", xlab="Banyak␣mobil␣yang␣terjual",
     main="Diagram␣Pencar") ## menggambar diagram pencar
lm(y~x) ## mencari koefisien-koefisien regresi
abline(lm(y~x)) ## menggambar garis regresi
```

◀

Distribusi, rata-rata, dan simpangan baku dari pendapatan untuk populasi secara berurut-urut disebut distribusi bersyarat, rata-rata bersyarat, dan simpangan baku bersyarat dari variabel respons "pendapatan" terhadap nilai variabel prediktor "banyak mobil yang terjual". Istilah-istilah tersebut digunakan secara

umum di sebarang variabel prediktor dan respons. Berikut adalah definisi distribusi, rata-rata, dan simpangan baku bersyarat:

Distribusi, rata-rata, dan simpangan baku bersyarat

Misalkan x adalah variabel prediktor dan y adalah variabel respons dalam suatu populasi. Misalkan juga x_p menyatakan suatu nilai tertentu dari variabel prediktor dan anggap subpopulasi yang terdiri dari seluruh anggota dari populasi yang nilai variabel prediktornya adalah x_p.

(i) **Distribusi bersyarat dari variabel respons yang berhubungan dengan** x_p adalah distribusi dari seluruh nilai yang mungkin dari variabel respons yang telah dijelaskan di subpopulasi.

(ii) **Rata-rata bersyarat dari variabel respons yang berhubungan dengan** x_p adalah rata-rata dari seluruh nilai yang mungkin dari variabel respons yang telah dijelaskan di subpopulasi.

(iii) **Simpangan baku bersyarat dari variabel respons yang berhubungan dengan** x_p adalah simpangan baku dari seluruh nilai yang mungkin dari variabel respons yang telah dijelaskan di subpopulasi.

Variabel respons adalah variabel yang akan diukur atau diamati, biasa juga dikenal dengan sebutan **variabel terikat**, sedangkan **variabel prediktor** adalah variabel yang digunakan untuk memprediksi atau menjelaskan variabel respons, dikenal juga dengan nama **variabel bebas**.

Dengan menggunakan terminologi tersebut, kita dapat menyatakan beberapa asumsi untuk melakukan inferensi dalam analisis regresi, yaitu:

(i) **Garis regresi populasi**. Terdapat dua konstanta β_0 dan β_1 sedemikian hingga setiap nilai x dari variabel prediktor, rata-rata bersyarat dari variabel responsnya adalah $\beta_0 + \beta_1 x$

(ii) **Simpangan baku yang sama**. Simpangan baku bersyarat dari variabel responsnya sama untuk seluruh nilai

variabel prediktor, yang dinyatakan dengan simbol umum simpangan baku, yaitu σ.

(iii) **Populasi normal**. Untuk setiap nilai variabel prediktor, distribusi bersyaratnya mengikuti distribusi normal.

(iv) **Pengamatan bebas**. Pengamatan variabel responsnya adalah saling bebas satu dengan yang lain.

3. Analisis residual

Metode untuk menguji asumsi 1-3 bergantung kepada analisis dari galat yang dilakukan dengan menggunakan persamaan regresi untuk memprediksi nilai variabel respons yang diamati, yaitu, selisih antara nilai variabel respons yang diamati dan nilai yang diprediksi. Selisih tersebut disebut sebagai **residual**, yang biasa dinyatakan sebagai **e**, sehingga:

$$\boxed{e_i = y_i - \hat{y}_i} \tag{8.5}$$

Gambar 8.3 menunjukkan bagaimana residual dari suatu titik dalam sebuah kurva regresi. Kita dapat menyatakan galat baku dari estimasi residualnya dengan rumus sebagai berikut:

$$\boxed{s_e = \sqrt{\frac{\text{JGK}}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}}} \tag{8.6}$$

JGK adalah jumlah galat kuadrat (*SSE, Error of Sum Squared*).

Kita dapat menunjukkan bahwa jumlah residualnya selalu nol, sehingga dapat disimpulkan bahwa $\bar{e} = 0$. Sebagai akibatnya, galat bakku dari estimasinya pada dasarnya sama dengan simpangan baku residual, sehingga galat baku dari estimasinya sering disebut sebagai **simpangan baku residual**.

Kita dapat menganalisis residual untuk memutuskan apakah asumsi (i)-(iii) untuk inferensi regresi terpenuhi karena asumsi tersebut dapat diterjemahkan ke dalam kondisi-kondisi residual.

Sehubungan dengan asumsi (i), titik-titik data harus tersebar pada garis regresi (sampel), yang berarti bahwa residual harus tersebar pada sumbu-x. Mengingat asumsi (ii), variasi nilai yang diamati dari variabel respons harus tetap mendekati konstan dari satu nilai variabel prediktor ke yang berikutnya, yang

Langkah-langkah menghitung galat baku estimasi	
Langkah 1	Buatlah tabel seperti berikut:

x	y	$\hat{y}$	e	e^2

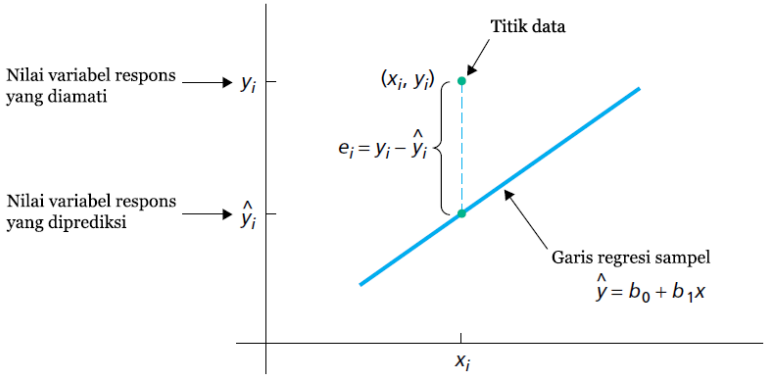
Langkah 2	Carilah nilai $\hat{y}$ untuk masing-masing x, kemudian tulis hasilnya di kolom $\hat{y}$
Langkah 3	Hitunglah nilai e dengan menggunakan rumus 8.5 kemudian tulis hasilnya di kolom e
Langkah 4	Kuadratkan nilai e, kemudian tulis hasilnya di kolom e^2
Langkah 5	Jumlahkan masing-masing kolom, tulis hasilnya di baris paling bawah
Langkah 6	Carilah s_e dengan menggunakan rumus 8.6

berarti residual harus jatuh kira-kira dalam garis horizontal. Dalam Asumsi (iii), untuk setiap nilai variabel prediktor, distribusi nilai observasi yang sesuai dari variabel respons harus berbentuk lonceng, yang menyiratkan bahwa garis horizontal harus dipusatkan dan simetris pada sumbu-x.

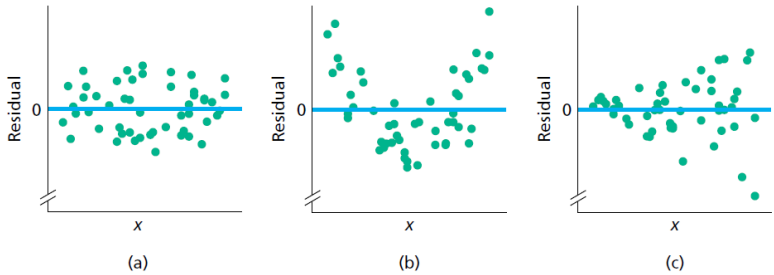
Kurva residual menggambarkan hubungan antara variabel prediktor pada sumbu-x dan residualnya pada sumbu-y. Jika titik-titik pada kurva residual terletak acak di sekitar garis horizontal, berarti suatu model regresi linear sesuai dengan datanya.

Selanjutnya, dengan mempertimbangkan keempat asumsi regresi secara bersamaan, kita dapat menganggap residual sebagai pengamatan independen dari variabel yang memiliki distribusi normal dengan rata-rata 0 dan standar deviasi σ. Jadi kurva probabilitas normal residual haruslah linear.

Suatu kurva dari residual terhadap nilai variabel prediktor yang diamati, yang disebut kurva residual, memberikan informasi yang kira-kira sama dengan sebaran titik data. Namun, sebuah kurva residual yang berpola pencar seperti kelengkungan dan standar deviasi tidak-konstan menjadi lebih mudah. Untuk



Gambar 8.3: Residual dari suatu titik



Gambar 8.4: Kurva residual yang menunjukkan pelanggaran asumsi

mengilustrasikan penggunaan kurva residual untuk diagnostik regresi, mari kita perhatikan tiga kurva pada Gambar 8.4.

Pada Gambar 8.4 (a), residual tersebar di sepanjang sumbu-x (residual = 0) dan jatuh di sepanjang garis horizontal sehingga asumsi (i) dan (ii) dipenuhi. Gambar 8.4 (b), kurva ini menunjukkan bahwa hubungan antara variabel-variabelnya membentuk suatu kurva sehingga asumsi (i) tidak dipenuhi. Sedangkan pada Gambar 8.4 (c), kurva menunjukkan bahwa simpangan baku kondisional bertambah seiring dengan pertambahan x, hal ini menunjukkan bahwa asumsi (ii) tidak dipenuhi.

Contoh 8.29. Buatlah kurva residual dari tabel Tabel 8.1 dan

hitunglah galat baku estimasinya.

Penyelesaian Dari contoh 5 telah kita peroleh bahwa persamaan regresinya adalah $\hat{y} = 0,396 + 0,106x$, sehingga kita dapat menghitung tabel residunya sebagai berikut:

Tabel 8.3: Tabel residual dari data penjualan mobil

Perusahaan	x	y	$\hat{y}$	e	e^2
F	8,5	1,5	1,297	0,203	0,041
E	13,4	1,4	1,816	-0,416	0,173
D	19,1	2,8	2,421	0,379	0,144
C	20,8	2,1	2,601	-0,501	0,251
B	29,0	3,9	3,470	0,43	0,185
A	63,0	7,0	7,074	-0,074	0,005
$\sum$	153,8	18,7	18,679	0,021	0,800

Dari tabel tersebut, kita dapat menghitung galat baku estimasinya, yaitu

$$s_e = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{0,800}{4}} = 0,224$$

.

Kurva residual diperoleh dengan menggambar variabel prediktor x terhadap nilai residual e, sehingga diperoleh:

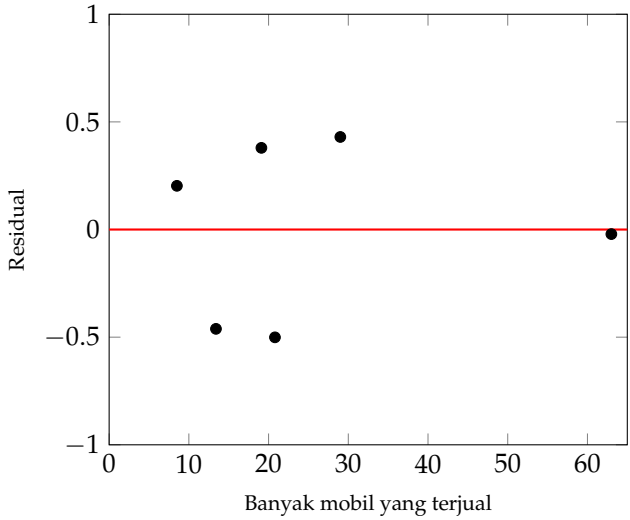
Source Code 8.2: *Syntax* R untuk membuat kurva residual

```
x<-c(63,29,20.8,19.1,13.4,8.5) ## nilai dari variabel bebas
y<-c(7,3.9,2.1,2.8,1.4,1.5) ## nilai dari variabel terikat
lm(y~x) ## mencari koefisien-koefisien
e = resid(lm(y~x)) ## mencari residual
plot(x,e, ylab="Residual", xlab="Banyak␣mobil␣yang␣terjual",
     main="Kurva␣Residual") ## menggambar kurva residual
abline(0, 0) ## menggambar garis horizontal
```

◀

4. Pengujian Hipotesis Koefisien Regresi

Misalkan variabel x dan y memenuhi asumsi inferensi regresi, maka untuk setiap nilai x, distibusi bersyarat dari variabel responsnya mengikuti distribusi normal dengan rata-rata $\beta_0 + \beta_1 x$



Gambar 8.5: Kurva residual

dan simpangan baku σ. Salah satu hal yang menarik untuk dibahas adalah koefisien regresi atau kemiringan regresi, β_1 perlu dicek apakah $\beta_1 = 0$. Jika nilai $\beta_1 = 0$, maka untuk setiap nilai variabel prediktor x, distribusi bersyarat dari variabel responsnya mengikuti distribusi normal dengan rata-rata β_0 (karena $\beta_0 + 0 \cdot x$) dan simpangan baku σ. Karena x tidak muncul pada dua parameter tersebut, penggunaan variabel y sebagai prediktor merupakan hal yang sia-sia saja. Maka dari itu, kita dapat memutuskan apakah x berguna sebagai prediktor (linear) dari y dengan melakukan uji hipotesis. Berikut ini adalah rumusan hipotesisnya:

$$\begin{aligned} \mathbf{H_0} &: \beta_1 = 0 \quad (x \text{ tidak berguna untuk memprediksi } y). \\ \mathbf{H_1} &: \beta_1 \neq 0 \quad (x \text{ berguna untuk memprediksi } y). \end{aligned}$$

Kita menguji hipotesis berdasarkan β_1 (kemiringan garis regresi populasi) dengan statistik b_1 (kemiringan garis regresi sampel). Untuk menjelaskan bagaimana metode ini bekerja, mari kita tinjau kembali data penjualan mobil di Amerika Serikat pada

tahun 2018.

Dengan banyak mobil yang terjual sebagai variabel prediktor dan pendapatan sebagai variabel respons, diperoleh persamaan regresinya yaitu $\hat{y} = 0,396 + 0,106x$, secara khusus, koefisien regresinya dari garis regresi sampel, $b_1 = 0,106$.

Sekarang kita anggap bahwa seluruh sampel dari enam perusahaan yang banyak mobil yang terjual sama dengan kolom kedua di Tabel 8.1. Untuk sampel tersebut, koefisiennya, b_1, dari garis regresi sampel berbeda dari satu sampel ke sampel lainnya . Distribusi ini disebut **distribusi sampling dari kemiringan garis regresi**. Dari asumsi untuk kesimpulan regresi, kita dapat menunjukkan bahwa distribusi ini adalah distribusi normal yang rata-ratanya adalah kemiringan, β_1, dari garis regresi populasi.

Distribusi *sampling* dari kemiringan garis regresi

Misalkan x dan y memenuhi asumsi inferensi regresi, sehingga untuk sampel sebanyak n, dengan nilai $x_1, x_2, \cdots, x_n$ yang sama untuk variabel prediktornya, sifat-sifat berikut berlaku untuk koefisien garis regresi sampel b_1

(i) Rata-rata b_1 sama dengan kemiringan garis regresi populasi, yang berarti $\mu_{b_1} = \beta_1$ (misalnya kemiringan garis regresi sampel adalah estimator takbias dari kemiringan garis regresi populasi).

(ii) Simpangan baku b_1 adalah $\sigma_{b_1} = \dfrac{\sigma}{\sqrt{s_{xx}}}$

(iii) Variabel b_1 berdistribusi normal

Sebagai akibatnya, variabel bakunya,

$$z = \frac{b_1 - \beta_1}{\sigma/\sqrt{s_{xx}}}$$

mengikuti distribusi normal. Tapi, variabel ini tidak dapat dijadikan basis untuk uji statistik yag dibutuhkan karena simpangan baku bersyarat yang umum, σ tidak diketahui sehingga kita mengganti σ dengan estimasi galat standar sampelnya, s_e, sehingga kita peroleh variabel baru yang berdistribusi t.

Misalkan variabel x dan y memenuhi keempat asumsi inferensi regresi. Untuk sampel sebanyak n, dengan nilai $x_1, x_2, \cdots, x_n$ yang

sama untuk variabel prediktornya,

$$t = \frac{b_1 - \beta_1}{s_e / \sqrt{s_{xx}}} \tag{8.7}$$

dengan derajat kebebasan (DK) = $n - 2$. Untuk pengujian hipotesis dengan $H_0 : \beta_1 = 0$, maka persamaan 8.7 menjadi

$$\boxed{t = \frac{b_1}{s_e / \sqrt{s_{xx}}}} \tag{8.8}$$

dimana persamaan 8.8 adalah uji statistik yang kita pergunakan.

Contoh 8.30. Ujilah signfikansi koefisien regresi dari data pada Tabel 8.1. Gunakan $\alpha = 0,05$.

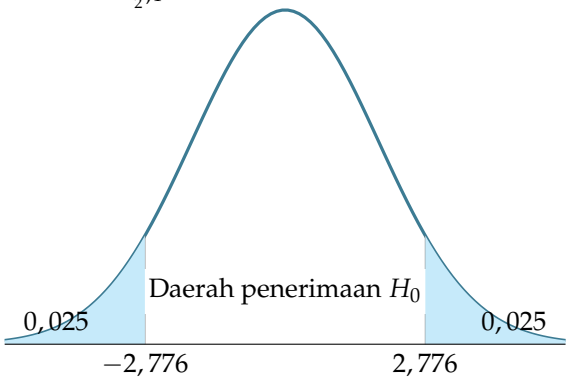
Penyelesaian Gunakan langkah-langkah uji hipotesis. ◀

Langkah 1 `Buatlah hipotesisnya.`

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Langkah 2 `Carilah nilai kritisnya dari tabel` t

Karena $\alpha = 0,05$ dan DK=4 maka diperoleh nilai kritis dari tabel $t_{\frac{\alpha}{2},4} = \pm 2,776$.



Langkah 3 `Hitung statistik ujinya.`

Dari contoh sebelumnya, diperoleh nilai-nilai $b_1 = 0,106$, $s_e = 0,224$, dan $s_{xx} = 1.916,853$.

diperoleh $t = \dfrac{b_1}{s_e/\sqrt{s_{xx}}} = \dfrac{0,106}{0,224/\sqrt{1.916,853}} = 20,783$

Langkah 4 `Buat keputusan berdasarkan statistik uji dan nilai kritis.`

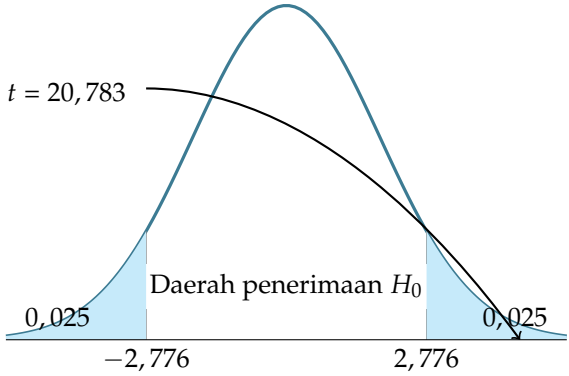
Tolak H_0 karena nilai uji berada pada daerah penolakan H_0.

5. Pengujian Hipotesis Konstanta Regresi

Selain pengujian hipotesis untuk koefisien regresi, kita juga dapat menguji hipotesis untuk konstanta regresi. Berikut ini adalah rumusan hipotesisnya:

Kita menguji hipotesis berdasarkan β_0 (titik potong garis regresi populasi) dengan statistik b_0 (titik potong garis regresi sampel). Untuk menjelaskan bagaimana metode ini bekerja, mari kita tinjau kembali data penjualan mobil di Amerika Serikat pada tahun 2018.

Dengan banyak mobil yang terjual sebagai variabel prediktor



Langkah 5 `Tariklah kesimpulan.`
Pada level signifikansi 5%, data tersebut memberikan cukup bukti untuk menyimpulkan bahwa koefisien garis singgungnya tidak sama dengan nol sehingga variabel banyaknya mobil berguna untuk memprediksi pendapatan.

$$\mathbf{H_0} : \beta_0 = 0$$
$$\mathbf{H_1} : \beta_0 \neq 0$$

dan pendapatan sebagai variabel respons, diperoleh persamaan regresinya yaitu $\hat{y} = 0,396 + 0,106x$, secara khusus, koefisien regresinya dari garis regresi sampel, $b_0 = 0,396$.

Misalkan variabel x dan y memenuhi keempat asumsi inferensi regresi. Untuk sampel sebanyak n, dengan nilai $x_1, x_2, \cdots, x_n$ yang sama untuk variabel prediktornya,

$$t = \frac{b_0 - \beta_0}{s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}} \tag{8.9}$$

dengan derajat kebebasan (DK) = $n - 2$. Untuk pengujian hipotesis

dengan $H_0 : \beta_0 = 0$, maka persamaan 8.9 menjadi

$$\boxed{t = \frac{b_0}{s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}}} \tag{8.10}$$

dimana persamaan 8.8 adalah uji statistik yang kita pergunakan.

Contoh 8.31. Ujilah signfikansi konstanta regresi dari data pada Tabel 8.1. Gunakan $\alpha = 0,05$.

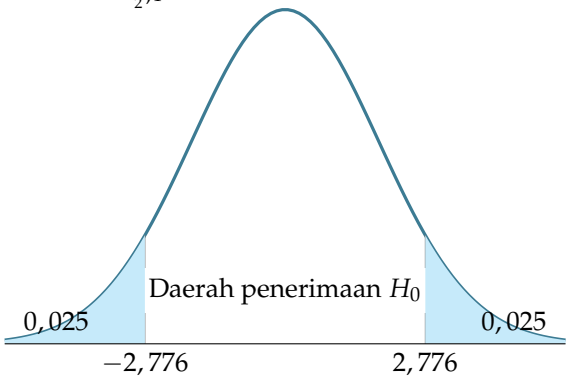
Penyelesaian Gunakan langkah-langkah uji hipotesis. ◀

Langkah 1 `Buatlah hipotesisnya.`

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

Langkah 2 `Carilah nilai kritisnya dari tabel` t

Karena $\alpha = 0,05$ dan DK=4 maka diperoleh nilai kritis dari tabel $t_{\frac{\alpha}{2},4} = \pm 2,776$.



6. Rangkuman

Setelah mendapatkan nilai korelasi maka tahapan berikutnya adalah memprediksi nilai y dari sebarang nilai x. Suatu teknik untuk memprediksi dinamakan sebagai regresi linear.

Untuk memprediksi maka diperlukan estimasi dari konstanta dan koefisien dari model regresi dengan metode least squared.

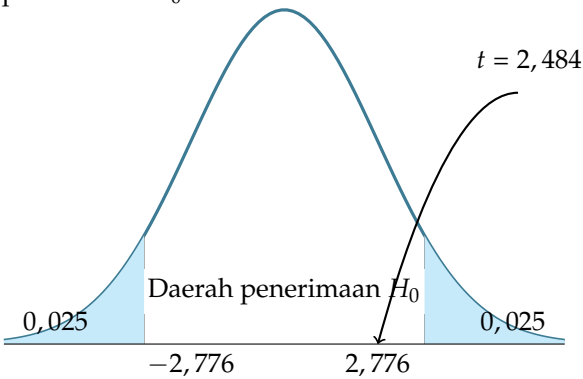
Langkah 3 `Hitung statistik ujinya.`

Dari contoh sebelumnya, diperoleh nilai-nilai $b_0 = 0,396, s_e = 0,224, \bar{x} = 25,63$, dan $s_{xx} = 1.916,853$. diperoleh:

$$t = \frac{b_0}{s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}} = \frac{0,396}{0,224\sqrt{\frac{1}{6} + \frac{(25,63)^2}{1.196,853}}} = 2,484$$

Langkah 4 `Buat keputusan berdasarkan statistik uji dan nilai kritis.`

Gagal menolak H_0 karena nilai uji berada pada daerah penerimaan H_0.



Langkah 5 `Tariklah kesimpulan.`

Pada level signifikansi 5%, data tersebut memberikan cukup bukti untuk menyimpulkan bahwa konstanta regresinya sama dengan nol.

Setelah model didapatkan, maka diperlukan evaluasi terhadap model tersebut yang disebut sebagai analisis residual.

Jika residual mengarah terhadap nilai nol maka ada indikasi bahwa model yang dihasilkan adalah baik dan sesuai dengan asumsi kenormalan

C. Perlatihan

Kerjakan soal-soal berikut ini untuk mempertajam pemahaman Saudara!

(1) Seorang peneliti ingin melihat adanya hubungan antara jumlah kasus cacar dan parotitis yang dilaporkan selama lima tahun terakhir ini di Amerika Serikat. Berikut ini adalah data yang berhasil dihimpun oleh peneliti tersebut

Kasus cacar	43	140	71	63	112
Kasus parotitis	800	454	1991	2612	370

(a). Tentukan persamaan regresinya, (b). Buatlah kurva residunya, (c). Ujilah apakah koefisien regresinya berguna untuk memprediksi variabel prediktornya

(2) Suatu sampel acak mengenai harapan hidup pria dan wanita dari negara dunia ketiga dipilih. Berikut ini adalah datanya:

D. Bahan Bacaan

Bluman, A. G. 2018. *Elementary Statistics: A Step By Step Approach (10th Edition)*. New York City: McGraw-Hill Education.

Johnson, R. A., Bhattacharyya, G. K. 2014. *Statistics: Principles and Methods (7th Edition)*. Hoboken: Wiley.

Walpole, M. 2013. *Probability and Statistics for Engineers and Scientist: Global Edition (9th Edition)*. Mumbai: Pearson.

Weiss, N. A. 2017. *Introductory Statistics: Global Edition (10th Edition)*. Harlow: Pearson.

Pria	59,7	72,9	41,9	46,2	50,3	43,2
Wanita	63,8	77,8	44,5	48,3	54,0	43,5

(a). Tentukan persamaan regresinya, (b). Buatlah kurva residunya, (c). Ujilah apakah koefisien regresinya berguna untuk memprediksi variabel prediktornya, (d). Lakukan uji hipotesis untuk konstanta regresinya.

(3) Seorang arsitek ingin menentukan hubungan antara ketinggan bangunan (dalam meter) dan banyak lantai di dalam bangunan. Dia telah menghimpun sampel dari delapan gedung di Kota Chicago seperti pada tabel berikut ini:

Jumlah lantai	64	68	50	48	32	46	58	45
Keting-gian	303	257	223	206	197	193	185	183

(a). Tentukan persamaan regresinya, (b). Buatlah kurva residunya, (c). Ujilah apakah koefisien regresinya berguna untuk memprediksi variabel prediktornya, (d). Lakukan uji hipotesis untuk konstanta regresinya.

BAB IX
UJI *CHI SQUARE*

A. Pendahuluan

1. Deskripsi Singkat

Teknik inferensi statistika yang telah dipelajari sejauh ini hanya sebatas pada uji hipotesis dan selang kepercayaan untuk parameter populasi, seperti rata-rata populasi dan proporsi populasi. Dalam bab ini, kita akan mempelajari tiga prosedur inferensi yang tidak berhubungan dengan parameter populasi. Tiga prosedur ini sering disebut sebagai prosedur *chi-square* karena prosedur-prosedur ini hanya bergantung pada distribusi *chi-square*.

2. Tujuan Pembelajaran

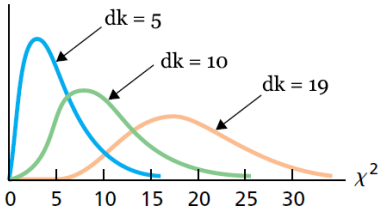
a. Menjelaskan pengetahuan tentang uji *chi-square* satu sampel
b. Menerapkan pengetahuan uji *chi-square* satu sampel dalam kehidupan sehari-hari.
c. Menjelaskan pengetahuan tentang uji *chi-square* dua sampel
d. Menerapkan pengetahuan uji *chi-square* dua sampel dalam kehidupan sehari-hari.

B. Materi

1. Distribusi *Chi-Square*

Prosedur inferensi statistik yang dibahas dalam bab ini bergantung pada sebuah distribusi *chi-square* (dibaca: kai skwer). Huruf *chi* merupakan huruf yunani χ.

Sebuah variabel akan berdistibusi *chi-square* jika distribusinya mempunyai bentuk kurva yang spesial, yaitu kurva condong-kanan–yang disebut kurva *chi-square* (χ^2). Terdapat sangat banyak distribusi *chi-square* (dan kurva (χ^2)) dengan derajat kebebasan



Gambar 9.1: Kurva χ^2 untuk beberapa derajat kebebasan

yang berbeda-beda. Gambar 9.1 berikut ini menunjukkan tiga derajat kebebasan kurva-χ^2 yang berbeda-beda dan beberapa sifat-sifatnya.

Sifat-sifat kurva χ^2	
(i)	Luas daerah di bawah kurva χ^2 sama dengan 1.
(ii)	Suatu kurva χ^2 dimulai dari 0 pada sumbu horizontal dan meluas tak hingga ke kanan mendekati namun tidak pernah menyentuh sumbu horizontal.
(iii)	Kurva χ^2 miring ke kanan.
(iv)	Seiring bertambahnya derajat kebebasan, kurva χ^2 nya terlihat mirip dengan kurva normal.

2. Uji Kebaikan Suai

Uji kebaikan suai adalah salah satu uji statistik yang banyak digunakan. Tujuan dari uji kebaikan suai ini adalah untuk membandingkan distribusi pengamatan dan distribusi harapan. Misalnya kita ingin melihat apakah ada perbedaan antara jumlah kriminalitas di suatu kota untuk empat jenis tindakan kriminal. Suatu sampel acak dari 160 pelaku ditunjukkan pada Tabel 9.1 berikut ini:

Karena frekuensi untuk setiap tipe kejahatan diperoleh dari sampel, frekuensi sebenarnya disebut **frekuensi pengamatan**. Frekuensi yang diperoleh dengan kalkulasi (jika tidak ada preferensi) disebut **frekuensi harapan**. Untuk menghitung frekuensi harapan, ada dua aturan yang harus diikuti

- Jika frekuensi harapannya sama, maka frekuensi harapannya

Tabel 9.1: Jenis kriminalitas

Kategori kejahatan	Pengamatan
Penjambretan	38
Pencurian	50
Penyalahgunaan NAPZA	28
Perampokan	44
Jumlah	160

E dapat dihitung dengan menggunakan rumus $E = n/k$, dimana n jumlah pengamatan dan k adalah jumlah kategori.

- Jika seluruh frekuensi harapannya tidak sama, maka frekuensi harapannya, E, dapat dihitung dengan rumus $E = n \cdot p$, dimana n jumlah pengamatan dan p adalah peluang untuk kategori yang dimaksud.

Uji kebaikan suai digunakan untuk mengklaim apakah suatu frekuensi pengamatan cocok dengan distribusi frekuensi harapan.

Jika melihat Tabel 9.1, jika tidak ada perbedaan, kita berharap ada $160 \div 4 = 40$ pelaku untuk setiap kategori. Tabel lengkapnya adalah sebagai berikut:

Tabel 9.2: Jenis kriminalitas dan frekuensinya

Kategori kejahatan	Pengamatan	Harapan
Penjambretan	38	40
Pencurian	50	40
Penyalahgunaan NAPZA	28	40
Perampokan	44	40
Jumlah	160	160

Frekuensi pengamatan hampir selalu berubah dengan frekuensi harapan akibat dari galat *sampling*, yaitu nilai-nilainya akan berubah dari sampel satu ke sampel yang lain. Timbul suatu pertanyaan baru: apakah perbedaan ini signifikan atau hanya

kebetulan saja? Uji kebaikan suai *chi-square* akan berguna dalam memperoleh jawabannya.

Sebelum melakukan pengujian, kita harus menentukan hipotesisnya. Hipotesis nol harus berupa pernyataan yang menandakan bahwa tidak ada perbedaan atau tidak ada perubahan. Berikut ini adalah rumusan hipotesisnya:

$\mathbf{H_0}$: Tidak ada perbedaan di antara tiap kategori
$\mathbf{H_1}$: Terdapat perbedaan di antara tiap kategori.

Selanjutnya kita memerlukan ukuran perbedaan antara frekuensi pengamatan, O, dan frekuensi harapan, E, sehingga statistik uji yang digunakan untuk uji kebaikan suai *chi-square* adalah

$$\boxed{\chi^2 = \sum \frac{(O-E)^2}{E}} \tag{9.1}$$

dengan derajat kebebasan $DK = n - 1$

Perhatikan bahwa nilai statistik ujinya berdasarkan perbedaan antara nilai pengamatan dan nilai harapan. Jika nilai yang diamati berbeda yang cukup signifikan nilai harapan, maka terdapat cukup bukti untuk menolak hipotesis nol. Perhatikan juga bahwa nilai χ^2 tak pernah negatif. Uji ini adalah uji satu arah (arah kanan) karena H_0: baik-suai dan H_1: tak baik-suai yang berarti bahwa nilai χ^2 akan kecil untuk kasus pertama dan besar yntuk kasus kedua. Beberapa asumsi untuk uji kebaikan suai *chi-square* adalah

(i) Data diperoleh dari sampel acak.

(ii) Frekuensi harapan untuk masing-masing kategori haruslah lebih dari 5.

Dalam menggunakan uji kebaikan suai *chi-square*, para statistikawan telah menentukan bahwa frekuensi harapannya paling tidak 5, seperti yang telah dinyatakan dalam asumsi di atas, alasannya adalah distribusi *chi-square* adalah distribusi kontinu sedangkan uji kebaikan suai adalah diskret. Namun, distribusi yang kontinu adalah pendekatan yang baik dan dapat digunakan ketika nilai harapan untuk masing-masing kategorinya paling tidak 5. Jika suatu frekuensi harapan dari suatu kategori kurang

Langkah-langkah uji kebaikan suai *chi-square*	
Langkah 1	Buatlah hipotesis.
Langkah 2	Carilah nilai kritisnya.
Langkah 3	Buatlah tabel lengkap untuk menghitung statistik ujinya

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$

Langkah 4	Ambillah keputusan berdasarkan hasil nilai uji dan nilai kritisnya. Terima H_0 jika $\chi_{\text{hitung}^2} < \chi_{\text{tabel}^2}$ Tolak H_0 jika $\chi_{\text{hitung}^2} > \chi_{\text{tabel}^2}$
Langkah 5	Tariklah kesimpulan.

dari 5, maka suatu kategori dapat dikombinasikan atau digabung dengan kategori lain sehingga frekuensi harapannya menjadi lebih dari 5.

Contoh 9.32. Frekuensi harapan sama. Apakah ada bukti yang cukup dari klaim bahwa jumlah pelaku kejahatan pada Tabel 9.1 untuk tiap-tiap kategori sama? Gunakan taraf nyata $\alpha = 0,05$.

Penyelesaian Gunakan langkah-langkah uji kebaikan suai untuk menyelesaikannya ◀

Langkah 1 `Buatlah hipotesisnya.`
H_0: Tidak ada perbedaan jumlah pelaku kejahatan untuk tiap kategori
H_1: Terdapat perbedaan jumlah pelaku kejahatan untuk tiap kategori

Langkah 2 `Carilah nilai kritisnya dari tabel` χ^2
Karena $\alpha = 0,05$ dan DK = 3 maka diperoleh nilai kritis dari tabel $\chi^2_{0,05;4} = 7,815$.

Langkah 3 `Buatlah tabel lengkap untuk menghitung statistik ujinya.`
Frekuensi harapan dihitung dengan rumus $E = n/k = 160/4 = 40$, sehingga:

Kategori kejahatan	O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Penjambretan	38	40	-2	4	0,1
Pencurian	50	40	10	100	2,5
Penyalahgunaan NAPZA	28	40	-12	144	3,6
Perampokan	44	40	4	16	0,4
Jumlah	160	160	0	264	6,6

Langkah 4 `Ambillah keputusan berdasarkan hasil nilai uji dan nilai kritisnya.`
Gagal menolak H_0 karena $\chi_{\text{hitung}^2} < \chi_{\text{tabel}^2}$, yaitu $6,6 < 7,815$

Langkah 5 `Tariklah kesimpulan.`
Pada level signifikansi 5%, data tersebut tidak ada bukti bahwa terdapat perbedaan jumlah pelaku kejahatan untuk tiap-tiap kategori.

Source Code 9.1: *Syntax* R untuk uji *chi-square* frekuensi harapan sama

```
data<-c(38,50,28,44) ##data observasi
chisq.test(data)
```

Contoh 9.33. Frekuensi harapan tidak sama. Asosiasi Pengelola Kesehatan Amerika Serikat (*American Health Administrator Association*/AHAA) melaporkan hasil survei nasional mengenai jumlah penduduk senior yang masuk rumah sakit pada tahun lalu. Hasilnya adalah sebagai berikut: empat puluh persen belum pernah sama sekali, tiga puluh persen dirawat sekali, dua puluh persen dirawat dua kali, dan sepuluh persen dirawat tiga atau empat kali.

Sebuah survei yang dilakukan pada 150 penduduk Perumahan Bartow di daerah Florida Tengah menunjukkan bahwa 55 penduduk belum pernah dirawat di rumah sakit selama tahun lalu, lima puluh penduduk pernah dirawat di rumah sakit sebanyak satu kali, 32 penduduk dirawat sebanyak dua kali, dan

sisanya mengaku pernah dirawat sebanyak tiga atau empat kali.

Dapatkah kita menyimpulkan bahwa survei yang dilakukan di Perumahan Bartow konsisten dengan informasi yang dilaporkan oleh AHAA? Gunakan taraf nyata $\alpha = 0,05$.

Penyelesaian Gunakan langkah-langkah uji kebaikan suai untuk menyelesaikannya ◀

Langkah 1 `Buatlah hipotesisnya.`
$\mathbf{H_0}$: Tidak ada perbedaan jumlah penduduk lokal dan nasional yang masuk rumah sakit.
$\mathbf{H_1}$: Terdapat perbedaan jumlah penduduk lokal dan nasional yang masuk rumah sakit

Langkah 2 `Carilah nilai kritisnya dari tabel` χ^2
Karena $\alpha = 0,05$ dan DK = 3 maka diperoleh nilai kritis dari tabel $\chi^2_{0,05;4} = 7,815$.

Langkah 3 `Buatlah tabel lengkap untuk menghitung statistik ujinya.`
Karena diketahui presentasi untuk masing-masing kategori, frekuensi harapan dihitung dengan rumus $E = n \cdot p = 150 \cdot p$, sehingga:

Jumlah dirawat RS (kali)	O	E	$O - E$	$(O - E)^2$	$\frac{(O-E)^2}{E}$
0	55	60	-5	25	0,4167
1	50	45	5	25	0,5556
2	32	30	2	4	0,1333
≥ 3	13	15	-2	4	0,2667
Jumlah	150	150	0	58	1,3723

Langkah 4 `Ambillah keputusan berdasarkan hasil nilai uji dan nilai kritisnya.`
Gagal menolak H_0 karena $\chi_{\text{hitung}^2} < \chi_{\text{tabel}^2}$, yaitu $1,3723 < 7,815$

Langkah 5 `Tariklah kesimpulan.`
Pada level signifikansi 5%, data tersebut tidak ada bukti bahwa terdapat perbedaan jumlah pelaku kejahatan untuk tiap-tiap kategori.

Source Code 9.2: *Syntax* R untuk uji *chi-square* frekuensi harapan tidak sama

```
obsv<-c(55,50,32,13) ## frekuensi observasi
expt<-c(0.4,0.3,0.2,0.1) ## peluang
chisq.test(x=obsv,p=expt)
```

3. Uji Independensi *Chi-Square*

Ketika data dapat ditabulasi dalam bentuk tabel distribusi frekuensi, beberapa tipe hipotesis dapat diuji dengan menggunakan uji *chi-square*. Uji indepensi *chi-square* dilakukan untuk menguji apakah dua variabel saling bebas satu dan yang lain. Berikut ini adalah rumusan hipotesisnya:

H$_0$: Variabel-variabelnya saling bebas satu dengan yang lain.
H$_1$: Variabel-varibelnya tidak saling bebas satu dengan yang lain.

Data-data untuk dua variabelnya ditabulasi dalam suatu **tabel kontingensi**. Variabel yang pertama disebut *variabel baris*, sedangkan variabel yang lain disebut *variabel kolom*. Tabel kontingensi disajikan sebagai berikut:

Tabel 9.3: Tabel kontingensi

	Kolom 1	Kolom 2	Kolom 3
Baris 1	$C_{1,1}$	$C_{2,1}$	$C_{3,1}$
Baris 2	$C_{1,2}$	$C_{2,2}$	$C_{3,2}$

Frekuensi yang diamati diperoleh dari data sampel. Frekuensi harapan diperoleh rumurs berikut:

$$\boxed{E = \frac{(\text{jumlah baris})\,(\text{jumlah kolom})}{\text{jumlah total}}} \tag{9.2}$$

dengan derajat kebebasan $DK = (\text{baris} - 1)\,(\text{kolom} - 1)$

Contoh 9.34. Suatu universitas melakukan survei terhadap 470 mahasiswanya. Pihak kampus ingin menilai apakah ada hubungan antara olahraga di kampus dan gaya hidup mahasiswa. Pilihan jawabannya adalah asrama, apartemen di dalam kampus, apartemen di luar kampus, dan di rumah. Data dari hasil surveinya ditunjukkan berikut ini:

Langkah-langkah uji independensi *chi-square*

Langkah 1 Buatlah hipotesis.

Langkah 2 Carilah nilai kritisnya.

Langkah 3 Buatlah tabel lengkap dan hitung statistik ujinya dengan menggunakan persamaan 9.2

	Kat. A		$\cdots$	Kat. B		dst	sum
	O	E		O	E		O
Kat. 1							
$\vdots$							
Kat. 2							
dst							
$\sum$							

Langkah 4 Ambillah keputusan berdasarkan hasil nilai statistik uji dan nilai kritisnya.

- ▶ Terima H_0 jika $\chi_{\text{hitung}^2} < \chi_{\text{tabel}^2}$
- ▶ Tolak H_0 jika $\chi_{\text{hitung}^2} > \chi_{\text{tabel}^2}$

Langkah 5 Simpulkan hasilnya.

Penyelesaian Gunakan langkah-langkah uji kebaikan suai untuk menyelesaikannya ◀

Source Code 9.3: *Syntax* R untuk uji independensi *chi-square*

```
asrama<-c(32,30,28)
apt_dalam_kampus<-c(74,64,42)
apt_luar_kampus<-c(110,25,15)
rumah<-c(39,6,5)
data=as.data.frame(rbind
                  (asrama,apt_dalam_kampus,apt_luar_kampus,rumah))
names(data) = c('tidak␣rutin', 'jarang','rutin') ## memberi nama kolom
chisq.test(data)
```

Tabel 9.4: Hasil survei dari universitas X

`Tempat tinggal mhs`	Tidak berolah-raga teratur	Jarang berolah-raga	Berolah-raga teratur	Total
Asrama	32	30	28	90
Aparte-men di dalam kampus	74	64	42	42
Aparte-men di luar kampus	110	25	15	150
Rumah	39	6	5	50
Total	255	125	90	332

Langkah 1 `Buatlah hipotesisnya.`
$\mathbf{H_0}$: Tempat tinggal mahasiswa saling bebas dengan keteraturan berolahraga.
$\mathbf{H_1}$: Tempat tinggal mahasiswa tidak saling bebas dengan keteraturan berolahraga.

Langkah 2 `Carilah nilai kritisnya dari tabel` χ^2
Karena $\alpha = 0,05$ dan $DK = (\text{baris} - 1)\,(\text{kolom} - 1) = (4-1)(3-1) = 6$ maka diperoleh nilai kritis dari tabel $\chi^2_{0,05;6} = 12,592$.

Langkah 3 `Buatlah tabel lengkap untuk menghitung statistik ujinya.`
Diperoleh $\chi^2 = 60,439$

Langkah 4 `Ambillah keputusan berdasarkan hasil nilai uji dan nilai kritisnya.`
Tolak H_0 karena $\chi_{\text{hitung}^2} > \chi_{\text{tabel}^2}$, yaitu $60,439 > 12,592$

4. Rangkuman

- **Uji kebaikan suai** digunakan untuk mengklaim apakah suatu frekuensi pengamatan cocok dengan distribusi frekuensi harapan.

Langkah 5 `Tariklah kesimpulan.`
Pada level signifikansi 5%, ada bukti kuat untuk menolak H_0 yang berarti bahwa tempat tinggal mahasiswa tidak saling bebas dengan keteraturan berolahraga.

C. Perlatihan

Kerjakan soal-soal berikut ini untuk mempertajam pemahaman Saudara!

Sebuah metode baru tengah dikenalkan di kalangan petugas kesehatan di suatu Rumah Sakit. Sebelum dilakukan secara serentak di level nasional, seorang peneliti melakukan survei di sebuah rumah sakit. Peneliti tersebut ingin mengetahui apakah respons dokter sama dengan respons para perawat. Berikut ini adalah hasil dari survei tersebut:

Grup	Prosedur baru	Prosedur lama	Abstain
Perawat	100	80	20
Dokter	50	120	30

Pada selang kepercayaan 95%, dapatkah disimpulkan bahwa preferensi perawat dan dokter terhadap prosedur sama?

D. Bahan Bacaan

Bluman, A. G. 2018. *Elementary Statistics: A Step By Step Approach (10th Edition)*. New York City: McGraw-Hill Education.

Johnson, R. A., Bhattacharyya, G. K. 2014. *Statistics: Principles and Methods (7th Edition)*. Hoboken: Wiley.

Walpole, M. 2013. *Probability and Statistics for Engineers and Scientist: Global Edition (9th Edition)*. Mumbai: Pearson.

Weiss, N. A. 2017. *Introductory Statistics: Global Edition (10th Edition)*. Harlow: Pearson.

GLOSARIUM