Maman A. Djauhari | Dyah E. Herwindiati

MEMBERSIHKAN DATA:
SATU VARIABEL

MEMBERSIHKAN DATA: SATU VARIABEL

Maman A. Djauhari | Dyah E. Herwindiati

Membersihkan Data: Satu Variabel

Penulis : Maman A. Djauhari & Dyah E. Herwindiati
Editor : Wiranto Handi
Penata Letak : Kiky Wulandari
Desain Cover : Kiky Wulandari

Edisi Asli Bhs. Indonesia

PENERBIT CAMPUSTAKA
Jl. H. Lebar No. 21B RT 02 RW 01
Meruya Selatan, Kembangan – Jakarta Barat 11650
Call/WA : +62 8888 03 11 30
E-mail : campustaka@gmail.com / info@campustaka.com
Website : www.campustaka.com

Djauhari, Maman A.
Herwindiati, Dyah E.

Membersihkan Data: Satu Variabel/Maman A. Djauhari, Dyah E. Herwindiati
– Jakarta: Campustaka, 2026
Anggota IKAPI No. 635/DKI/2024
1 jil, 14.8 x 21 cm, 222 hal
ISBN: 978-634-96395-7-6
1. Komputer
2. Membersihkan Data: Satu Variabel
I. Judul
II. Penulis

"*Fikiran adalah fakulti manusia yang memungkinkan seseorang mampu melihat tanda-tanda yang tersirat akan adanya pelajaran*."

Q.S. 45:13

Buku ini
Kami dedikasikan untuk

KECEMERLANGAN
IBU PERTIWI

(Kepada siapa kami banyak berhutang)

(*Center Outward Ordering*)

"Membersihkan data
tak ubahnya memisahkan
kacang-kacang buruk dari kelompok kacang baik
seperti diajarkan nenek moyang"

BUKU LAIN KARYA PARA PENULIS
(terbit dalam 10 tahun terakhir)

1. "Advanced monitoring techniques for complex process variability in manufacturing industry" UPM Press, 2016.
2. "Forecasting Methods: The needs of policymakers" UPM Press, 2018.
3. "Reliable Shewhart-type control charts for multivariate process variability" Lambert Academic Publishing, printed in Singapore, 2018. Tahun 2020 buku tersebut diterbitkan ulang di 8 negara Eropa dalam bahasa setempat;

    1) "Betrouwbare Shewhart-type controle diagrammen voor multivariate procesvariabiliteit" Uitgeverij Onze Kennis, 2020 (Belanda).
    2) "Grafici di controllo affidabili di tipo Shewhart-type per la variabilità di processo multivariate" Edizioni Sapienza, 2020 (Italia).
    3) "Zuverlässige Shewhart-kontrollkarten für multivariate prozessvariabilität" Verlag Unser Wissen, 2020 (Jerman).
    4) "Cartes de contrôle fiables de type Shewhart pour la variabilité des processus à plusieurs variables" Editions Notre Savoir, 2020 (Perancis).
    5) "Niezawodne wykresy kontrolne typu Shewhart-type control charts dla wielowymiarowej zmienności procesu" Wydawnictwo Nasza Wiedza, 2020 (Polandia).
    6) "Gráficos de controlo fiáveis do tipo Shewhart-type para variabilidade multivariada do processo" Edições Nosso conhecimento, 2020 (Portugis).
    7) "Надежные диаграммы управления типа Shewhart-для многомерной переменной процесса" Sciencia Scripts, 2020 (Rusia).
    8) "Gráficos de control fiables de tipo Shewhart para la variabilidad multivariante de los procesos" Ediciones Nuestro Conocimiento, 2020 (Spanyol).

4. "Indonesian Konjak: Its benefit in industry and food security" Scholar Press, Printed in Singapore, 2019.
5. "Ukuran Sampel: Formula generik bagi praktisi sains sosial" ITB Press, 2020.

6. “Teknik Memuaskan Editor & Reviewer Jurnal Internasional Berkelas” ITB Press, 2021.
7. “Teknik Membersihkan Data: Awas ... outliers ...Outlier ... OUTLIER!” ITB Press, 2021.
8. “Kontrol Kualitas Proses Kompleks” ITB Press, 2022.

PENGANTAR

"*It always seems impossible until it's done*."

Credited to Nelson Mandela

Hingga awal dekade1960an, tatkala orang berbicara tentang data, maka dapat dipastikan bahwa yang dimaksud adalah data statistik (*statistical data*). Artinya, data yang diproduksi melalui penggunaan metode statistik dan bentuknya numerik (Djauhari, 2020). Seiring dengan perkembangan dalam bidang ilmu komputer dan ilmu statistika, sejak awal 1960an pengertian data meliputi fakta tanpa makna yang bisa berupa gambar (*image*), foto, numerik, alfanumerik, simbol, karakter, suara, dll.

Perkembangan bidang ilmu komputer telah memungkinkan orang mengolah data non-numerik, terutama sejak teknologi internet menjadi terjangkau semua orang. Di lain pihak, ilmu statistika yang semula hanya terfokus pada analisis data konfirmatif (*confirmatory data analysis* atau CDA) dengan basisnya adalah teori probabilitas, sejak dekade itu orang dimungkinkan melakukan analisis data eksploratif (*exploratory data analysis* atau EDA).

Setidaknya ada tiga tokoh yang bertanggung jawab terhadap kemajuan ilmu statistika yang berujung pada EDA. Mereka berasal dari tiga benua yang berbeda dan bekerja secara independent; satu sama lain. Mereka tidak pernah berhubungan. Pertama adalah Chikio Hayashi dari Jepang di benua Asia yang memulai membangun fondasi matematis untuk menganalisis data kualitatif (Hayashi, 1951, dan Ohsume, 2004). Kedua, John W. Tukey dari benua Amerika yang memperkenalkan analisis data eksploratif khusus untuk data univariat, artinya hanya satu buah variable yang terlibat (Tukey, 1977). Dan yang ketiga adalah Jean-Paul Benzécri dari Perancis di benua Eropa yang membuat terobosan dan berhasil menerangi dunia statistik dengan apa yang dia sebut *Analyse des Correspondances* atau Analisis Korespondensi (Benzécri, 1960, Caillez dan Pages, 1976). Analisis ini digunakan untuk mengekplorasi informasi yang

tersembunyi di dalam tumpukan data bivariat (dua buah variabel yang saling berkorelasi) di mana kedua variabel tersebut berupa variabel kualitatif.

1. Evolusi Data

Berbagai kemajuan seperti tersebut di atas telah mengubah persepsi orang tentang data. Terutama sejak lahirnya Sains Data (*Data Science*). Pada saat ini, data dapat diartikan apa saja (tidak harus data statistik) baik itu gambar (*image*), foto, angka numerik, alfanumerik, simbol, karakter, musik, film, dll. Bahkan, sejak memasuki era media sosial, data berevolusi menjadi komoditas ekonomi dan informasi menjadi alat kekuasaan (*power*).

Sekilas tentang kelahiran Sains Data. Ada satu tokoh sentral yang menjembatani kemajuan ilmu statistika dan kemajuan ilmu komputer. Tokoh itu namanya Yves Escoufier juga dari Perancis yang pada awal 1960an membangun sistem komputasi untuk analisis data multivariat (banyak variabel yang saling berkorelasi) dengan hanya menggunakan komputer yang kemampuannya sedikit di atas kalkulator. Sekitar 30 tahun kemudian, karyanya itu melahirkan EMDA (*exploratory multivariate data analysis*). Inilah cikal bakal Sains Data.

Istilah "*Data Science*" belum pernah terdengar di komunitas saintifik internasional sebelum tahun 1992. Istilah itu muncul secara resmi untuk pertama kalinya dalam seminar bersama Prancis-Jepang yang kedua di Université de Montpellier 2 (UM2) pada bulan September tahun 1992. Yang mengumandangkannya adalah Yves Escoufier bersama Chikio Hayashi dkk. Prosiding seminar tersebut menjadi buktinya (Escoufier, et al., 1995). Sembilan tahun kemudian yakni tahun 2001, istilah itu populer di benua Amerika setelah terbit publikasi Cleveland (2001). Padahal tiga tahun sebelumnya, Hayashi (1998) telah berbicara tentang apa itu *Data Science*.

Kelahiran Sains Data telah menimbulkan dampak yang sangat hebat pada kehidupan politik, ekonomi, sosial, dan budaya. Seiring dengan itu, persepsi orang tentang data pun berkembang. Orang awam pun sekarang terbiasa membeli data. Namun, data numerik atau data yang telah dinumerikkan (identik dengan data statistik) tetap menjadi

domain yang didominasi oleh para ilmuwan. Data inilah yang menjadi pokok bahasan buku ini.

2. Data Harus Bersih

"Jangan mengolah sesuatu sebelum yakin bahwa sesuatu itu layak diolah." Begitulah salah satu pelajaran mendasar tentang kehidupan, yakni pelajaran berintegritas. Begitu pula, dalam menegakkan integritas keilmuan, hindari mengolah data kalau belum yakin data itu bersih. Dan memang, dalam praktik, pembersihan data adalah bagian yang paling banyak menyita waktu dan tenaga.

Membersihkan data tidak berbeda dengan menyingkap sebuah misteri. Begitu pula waktu dan tenaga yang dicurahkan untuk membersihkan data tidak berbeda dengan waktu dan tenaga seorang detektif profesional dalam menyingkap sebuah misteri.

Di dalam bukunya "The Adventure of the Copper Beeches," Arthur Conan Doyle berteriak "Data! Data! Data! I can't make bricks without clay." Doyle faham betul bahwa dalam memecahkan misteri diperlukan kemahiran hermeneutika dan logika. Secara logika, tidak mungkin membuat batu-bata tanpa tanah liat. Dan, sudah barang tentu, pemahaman yang tajam akan data tidak mungkin diperoleh tanpa kemampuan hermeneutika yang baik dalam menginterpretasikan data. Data yang dimaksud Doyle adalah fakta tanpa makna yang bersifat umum dan bisa berbentuk apa saja. Namun, sekali lagi, dalam buku ini perhatian akan difokuskan kepada data statistik; artinya, data dalam bentuk numerik atau data yang telah dinumerikkan.

Sherlock Holmes, sang pelakon utama dalam buku Doyle itu, mengalami keadaan frustrasi tatkala fakta yang ada kurang lengkap sehingga pemahamannya tentang misteri yang dihadapi tidak utuh. Sebagai akibatnya, proses pengambilan keputusan tidak mungkin berjalan dengan benar. Untuk itulah, Holmes membutuhkan data yang benar-benar bersih dan akurat tanpa cacat. Dia memerlukan informasi yang berkualitas.

Begitu pulalah para peneliti yang bekerja berbasiskan data. Mereka tak ubahnya seperti Sherlock Holmes; melakukan eksplorasi dan

investigasi untuk menggali dan mengais informasi yang tersembunyi di dalam tumpukan data dengan tujuan dapat tiba kepada keputusan yang sahih (*justified*) dan signifikan. Untuk itu, mereka membutuhkan data yang benar-benar bersih dan terpercaya. Itulah sebabnya, proses membersihkan data merupakan proses awal yang niscaya dilakukan sebelum data tersebut diolah dan dianalisis. Data yang bersih, akurat dan tepat waktu menjadi dasar dari berbagai pengambilan keputusan penting. Selain itu, untuk kegunaan analisis inferensial, sumber data seperti populasi dan distribusinya harus bisa dikenalpasti.

Tidak berbeda dengan Sherlock Holmes, para peneliti adalah para pencari fakta, pengumpul data, penyelidik, penyidik, dan sekaligus jaksa merangkap hakim dalam memutuskan suatu perkara (hipotesis). Sebagai contoh, apakah sebuah atau beberapa buah data yang bernilai ekstrim (*extreme values* atau *erroneous values*) sahih diputuskan sebagai data asing (*outlier*)? Di sini yang dimaksud data asing, atau *outlier* dalam bahasa statistik, adalah data yang berasal dari populasi yang berbeda dengan populasi kelompok data lainnya?

Para peneliti itu sadar dan faham bahwa proses memisahkan data asing dimulai dari proses mengidentifikasi calon *outlier*, dilanjutkan dengan proses mengenalpasti *outlier* dan diakhiri dengan proses konfirmasi. Dengan kata lain, dimulai dengan proses penyeldikan untuk mengidentifikasi calon tersangka *outlier*, lalu diikuti dengan proses penyidikan untuk mendeteksi apakah calon tersangka layak dijadikan tersangka *outlier* dan diakhiri dengan proses pengadilan atau penghakiman atau pengujian hipotesis apakah tersangka memang sah diputuskan sebagai *outlier*.

3. Proses Komputasi

Setiap analisis data dan/atau analisis statistik memerlukan proses komputasi yang cepat, akurat, murah dan terperinci. Selain itu, ia memerlukan pula proses visualisasi data. Dalam buku ini, semua proses komputasi dilaksanakan dengan bantuan Microsoft Excel (disingkat MS Excel). Penggunaan MS Excel didasarkan kepada falsafah "tidak membunuh nyamuk dengan bedil." MS Excel sudah lebih dari cukup...! Bahkan dengan MS Excel, para guru dan murid-muridnya tidak akan kesulitan belajar membersihkan data secara

mudah dan interaktif. Begitu pula dengan para pamong, polisi, tentara dan para pejabat berbagai agen BIG (*business, industry, government*) di seluruh pelosok nusantara.

Sedangkan untuk visualisasi data, semua gambar grafik dibuat dengan menggunakan perangkat lunak MINITAB. Perangkat lunak ini dipilih karena mampu menyajikan gambar berwarna dengan lebih jelas ketimbang MS Excel selain murah dan mudah diperoleh.

4. Catatan Penutup

Buku ini merupakan pengembangan dan pendalaman dari buku berjudul "Teknik membersihkan data: Awas ... outlier ... Outlier ... OUTLIER!" yang diterbitkan oleh ITB Press pada tahun 2021 dengan ISBN: 978-623-297-164-6 (Djauhari, 2021). Pengembangan dan pendalaman yang dimaksud meliputi pengorganisasian gagasan (*idea*), perumusan yang bersifat redaksional maupun pendalaman substansi.

Di dalam buku ini kami sajikan 13 (tigabelas) teknik/metode baru (*novelty*) yang merupakan hasil penelitian kami sendiri. Semuanya tersebar di lima bab yakni Bab 6 ada 1, Bab 7 ada 1, Bab 9 ada 2, Bab 11 ada 6, dan Bab 13 ada 3. Rinciannya sebagai berikut.

No,	Bab	Kebaruan
1	6	Konstanta pengali pada teknik Tukey
2	7	Konstanta pengali pada teknik Iglewicz-Hoaglin
3	9	Teknik IESD
4	9	Distribusi statistik penguji yang eksak dan nilai kritis eksak
5	11	Enam buah tabel nilai titik kritis (Tabel 11.1 – Tabel 11.6) untuk berbagai ukuran sampel n, berbagai nilai k (banyaknya *outlier* di dalam kelompok data), dan berbagai tingkat signifikansi
6	13	Teknik FMV
7	13	Metode MVV (*minimum vector variance*) yang kami singgung tatkala memperkenalkan teknik FMV

8	13	Satu tabel nilai titik kritis (Table 13.5) untuk *robust Mahalanibis squared distance* (RMSD) pada berbagai ukuran sampel n mulai dari n = 10 sampai dengan n = tak hingga, dan pada berbagai tingkat signifikansi 1%, 2,5%, 5% dan

Di samping 13 teknik/metode baru itu, seluruh isi Bab 15 yang terdiri atas 6 (enam) tabel data adalah bagian penelitian kami tatkala menyusun buku "*Forecasting Methods: The needs of policymakers*" bersama Lee Siaw Li. Buku ini dididterbitkan oleh UPM Press tahun 2018.

Akhir kata, petuah bijak dari para statistisi berikut ini patut dijadikan pedoman: "Tanganilah *outlier* dengan sangat hati-hati terutama kalau kita tidak mengetahui riwayatnya! Dan, jangan pula lupa untuk menggunakan beberapa teknik dalam penyelidikan calon tersangka *outlier*, penyidikan atau peningkatan status menjadi tersangka *outlier*, dan penghakiman atau pengujian hipotesis apakah tersangka sah sebagai *outlier*. Lalu, kombinasikanlah hasil teknik-teknik itu karena tidak ada yang namanya teknik terbaik."

Selamat berpetualang di medan intelektual yang gelap dan menantang...!

Salam dari kami para Penulis.

DAFTAR ISI

BAB 1. PENDAHULUAN: PERALATAN PALING MENDASAR

"*Docendo discimus — the best way to learn is by teaching*"

Latin principle

Dalam ilmu statistika ada kesepakatan bahwa sekelompok data dikatakan bersih apabila kelompok tersebut terbebas dari kehadiran data anomali dan/atau data asing (*outlier*). Yang dimaksud data anomali adalah data hilang (*missing data*) atau data yang nilainya tidak otentik (bukan yang seharusnya) sebagai akibat yang disebabkan oleh, misalnya, adanya anomali pada sistem pencatatan data. Data seperti ini perlu diperbaiki jika memungkinkan atau diasingkan jika tidak memungkinkan untuk diperbaiki.

Adapun yang dimaksud dengan data *outlier*, yang secara umum ditandai dengan nilainya yang ekstrim (*extreme values* atau *erroneous values*), adalah data yang bersumber dari populasi yang berbeda dengan populasi yang dikaji. Nah, data seperti ini perlu dipisahkan (jangan dibuang begitu saja tanpa argument saintifik) dan dianalisis tersendiri. Data ini tidak dilibatkan dalam analisis kelompok besar data lainnya. Mengapa? Karena keterlibatannya atau kehadirannya dapat mengganggu pola perilaku distributional kelompok besar data lainnya. Sebagai akibatnya, maka keputusan yang diambil akan melenceng dari yang seharusnya.

Oleh karena itulah, sebelum sekelompok data diolah dan kemudian dianalisis, pastikan tidak ada data anomali maupun *outlier*. Mengingat bahwa penanganan data anomali hanyalah bersifat teknis dan tidak bersifat probabilistis, maka seluruh pokok bahasan dalam buku ini difokuskan pada teknik memisahkan data *outlier* dari kelompok data lainnya.

1.1. Ilustrasi Sederhana

Teknik memisahkan outlier dari kelompok data lainnya dapat diilustrasikan secara sederhana dengan bantuan Gambar 1.1 berikut.

Gambar 1.1. Memisahkan kacang jelek dari kelompok kacang baik

Gambar ini memperlihatkan cara kerja teknologi klasik warisan nenek moyang kita tatkala memisahkan beberapa kacang yang jelek dari kelompok kacang yang baik. Teknologi nenek moyang terebut mengilustrasikan dengan baik dan sederhana bagaimana *outlier(s)* dipisahkan dari kelompok data lainnya. Kacang mewakili data, sedangkan media cawan dan air diganti dengan media matematik dan/atau media statistik. Kacang yang "mengambang" di permukaan air mengilustrasikan data *outlier(s)*. Sedangkan kacang yang "tenggelam" di dasar cawan adalah kelompok kacang/data yang terhindar dari *outlier(s)* dan akan diproses lebih lanjut untuk dianalisis. Bagaimana dengan kacang/data yang tidak mengambang di permukaan namun tidak tenggelam di dasar cawan? Di sinilah letak tantangan terberat dalam mengembangkan teknik separasi data *outlier*. Tantangan ini pula yang menjadi salah satu topik yang amat menantang (*hot topic*) dalam buku ini.

1.2. Keutamaan Membersihkan Data

Proses membersihkan data dari kehadiran *outlier* merupakan masalah terbesar dari seluruh rangkaian proses analisis data dan analisis

statistik. Para ahli statistika memperkirakan sekitar 80% aktivitas analisis statistik (dari mulai pengumpulan data, penyiapan data, pembersihan data, pengolahan data, analisis data, analisis statistikal sampai dengan interpretasi hasil analisis) adalah untuk memastikan bahwa data yang hendak dianalisis adalah data yang sehat dan bersih dari *outlier*.

Dari seluruh rangkaian aktivitas riset berbasis data statistik, aktivitas membersihkan data (*data cleansing*) berada pada urutan kedelapan. Semuanya ada 17 aktivitas berbeda dalam riset tersebut yang dimulai dari pengembangan ide riset (*research idea*) sampai dengan implementasi hasil riset di lapangan (*field action*). Ketujuhbelas aktivitas itu adalah,

1. Pengembangan ide riset (*Research idea*)
2. Tujuan riset (*Research goals*)
3. Masalah riset (*Research problems*)
4. Rancangan riset (*Research design*)
5. Rancangan pengambilan sampel (*Sampling design*)
6. Pengumpulan data (*Data collection*)
7. Penyiapan data (*Data preparation*)
8. Pembersihan data (*Data cleansing*)
9. Pengolahan data (*Data processing*)
10. Analisis data/analisis statistik (*Data analysis/statistical analysis*)
11. Dokumentasi hasil riset (*Research results*)
12. Interpretasi hasil riset (*Interpretation of research results*)
13. Penemuan dalam bentuk informasi baru (*New information*)
14. Penemuan dalam bentuk ilmu baru (*New scientific knowledge*)
15. Publikasi sebagai justifikasi ilmu baru (*Peer reviewed publication*)
16. Pengembangan kebijakan (*Policy development*)
17. Implementasi hasil riset di lapangan (*Field Action*)

Begitu utamanya kesehatan dan kebersihan data dalam riset. Sekali lagi, dari 17 aktivitas itu, aktivitas nomor delapan (proses pembersihan data) menyita sekitar 80% volume pekerjaan. Hanya data yang sehat dan bersih, ditambah dengan teknik analisis data yang

tepat, yang akan menghantarkan para peneliti kepada rumusan kebijakan dan tindakan lapangan yang tepat.

Proses pembersihan data merupakan fase yang amat penting dalam pemrosesan data setelah pengumpulan data. Fase ini melibatkan identifikasi dan koreksi kesalahan dalam sekumpulan data, seperti penanganan data yang hilang, penghapusan redundansi, dan penanganan outlier. Fase ini bertujuan untuk memastikan bahwa analisi data dan/atau analisis statistik akan memberikan hasil yang akurat dan realistis. Fase ini benar-benar harus dipahami dengan baik oleh para peneliti dan praktisi pengguna statistik, pembelajaran mesin (*machine learning*), sains data, dan penambangan data (*data mining*).

Data yang disebut outlier sering ditemukan di hampir setiap tumpukan data. Sebagian ahli mengasumsikan kehadiran outlier sebagai akibat kesalahan atau gangguan (noise) sistem pembangkit data. Dampak kehadirannya tidak dapat dielakkan; estimasi parameter akan bias dan berujung pada keputusan yang salah. Dalam pembelajaran mesin, misalnya, outlier dapat menyebabkan overfitting. Oleh karena itulah mengapa studi tentang outlier terus berkembang dan menjadi topik hangat dan menawarkan banyak peluang untuk eksplorasi maupun pembelajaran lebih lanjut. Bagi para pembaca yang berminat mendalami topik ini, kami persilahkan untuk mempelajari Han, et al. (2012) dan García, et al. (2015).

1.3. Peralatan Dasar

Setelah data benar-benar bersih dari kehadiran data *outlier(s)*, langkah berikutnya adalah pengolahan data untuk mempersiapkan analisis data eksploratif (EDA) dan analisis statistik inferensial. Pada langkah ini senjata utama yang diperlukan adalah "*The four plots*" atau kami sebut "Diagram-4" yang terdiri atas empat buah diagram berikut.

1. Diagram deretan data (*Run Sequence Plot* atau disebut juga *Run Chart*)
2. Diagram Lag-1 (*Lag-1 Plot*)
3. Histogram
4. Diagram probabilitas normal (*Normal Probability Plot*)

Setiap diagram tersebut memiliki peran masing-masing yang berbeda satu sama lain.

1. Diagram deretan data sesuai urutan observasi (*Run Chart*) adalah diagram pencar (*scatter plot*) antara nomor urut observasi (sumbu horizontal) dan nilai data (sumbu vertikal) yang dilengkapi dengan garis penghubung antara 2 titik data yang berturutan. Diagram ini diperlukan untuk mendapatkan gambaran tentang parameter lokasi data (*central tendency*) dan parameter variasi data (*variance/dispersion/spread*). Pada dasarnya,
    1) Diagram yang tampak mendatar (*flat*) dan tidak menunjukkan pergeseran pola sepanjang sumbu horizontal adalah pertanda bahwa lokasi data konstan sepanjang sumbu tersebut.
    2) Jika dispersi pada sumbu vertikal tidak banyak berbeda sepanjang sumbu horizontal, maka ini adalah pertanda bahwa variansi konstan sepanjang sumbu tersebut.
2. Diagram Lag-1 adalah diagram pencar antara data ke-i dan data ke-(i+1) dengan i = 1, 2, …, (n-1). Di sini n adalah banyaknya data (atau ukuran sampel). Diagram ini digunakan untuk menggambarkan tingkat kerandoman data (randomness). Jika diagram ini tidak menunjukkan pola tertentu (*structureless*), maka kita dapat mengasumsikan bahwa data bersifat acak (*random*).
3. Histogram adalah diagram yang dapat digunakan antara lain untuk memperkirakan kesimetrisan distribusi data. Jika histogram berbentuk seperti lonceng (*bell-shaped*), maka kita dapat mengasumsikan data berasal dari populasi berdistribusi simetris bahkan mungkin berdistribusi normal.
4. Diagram probabilitas normal adalah diagram pencar antara kuantil data dan kuantil distribusi normal. Pola yang tampak pada diagram ini menandakan sejauh mana asumsi kenormalan data dapat diterima. Jika diagram tersebut berpola garis lurus (*linear*), maka kita dapat mengasumsikan bahwa data berasal dari populasi berdistribusi normal.

Bagaimana cara kerja Diagram-4 itu? Berikut penjelasannya dengan contoh. Perhatikan 100 buah data acak pada Tabel 1.1 di Bab 1.

Diagram-4 untuk data tersebut disajikan pada Gambar 1.2. Pada gambar ini, (A), (B), (C) dan (D) berturut-turut menyatakan Diagram deretan data, Diagram Lag-1, Histogram, dan Diagram probabilitas normal. Adapun C1, C2 dan C3 pada gambar itu berturut-turut adalah (1) nomor data, (2) nilai data dari pertama sampai dengan keseratus, dan (3) nilai data Lag-1 dari data kedua sampai dengan keseratus. Nah, kalau keempat diagram itu kita perhatikan dengan seksama, tampak bahwa:

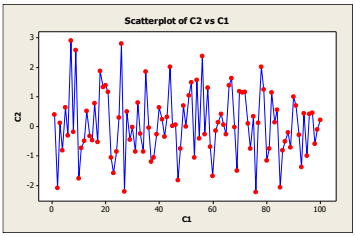
1) Diagram deretan data (A) mengindikasikan kesahihan asumsi bahwa lokasi konstan dan variasi data juga konstan.
2) Diagram Lag-1 (B) mengindikasikan bahwa data bersifat random.
3) Histogram (C) berbentuk hampir seperti lonceng mengindikasikan data berdistribusi simetris.
4) Diagram probabilitas normal (D) berpola linear mengindikasikan data berasal dari distribusi normal.

Penampakan Diagram-4 ini tidak mengherankan karena 100 data pada Tabel 1.1 di Bab 1 itu diperoleh melalui simulasi dari populasi distribusi normal standar (mean 0 dan deviasi standar 1) yang biasa ditulis N(0,1).

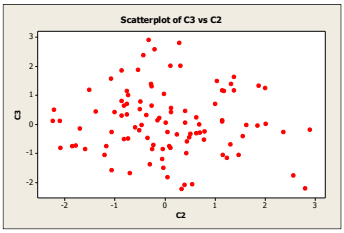
Tabel 1.1. Seratus buah data acak

No.	Data	No.	Data	No.	Data
1	0,4002	36	-0,0471	71	1,1483
2	-2,0913	37	-1,1982	72	1,1499
3	0,1116	38	-1,0631	73	0,0947
4	-0,8130	39	-0,2645	74	-0,7543
5	0,6259	40	0,6354	75	0,3360
6	-0,3239	41	0,2384	76	-2,2280
7	2,8906	42	-0,3634	77	0,1175
8	-0,2000	43	0,3136	78	2,0020
9	2,5671	44	2,0100	79	1,2362
10	-1,7675	45	0,0151	80	-1,1587
11	-0,7336	46	0,0571	81	-0,7577
12	-0,4894	47	-1,8386	82	1,1455
13	0,5075	48	-0,7499	83	0,1224

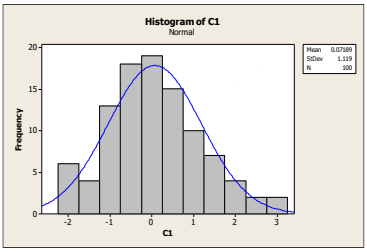
14	-0,3373	49	0,6904	84	0,5465
15	-0,4795	50	-0,0153	85	-2,0766
16	0,7799	51	1,0363	86	-0,8107
17	-0,5298	52	1,4778	87	-0,5119
18	1,8675	53	-1,0639	88	-0,2217
19	1,3266	54	1,5617	89	-0,7253
20	1,3833	55	-0,4221	90	1,0074
21	1,1624	56	2,3668	91	0,7025
22	-1,0600	57	-0,2665	92	-0,2926
23	-1,5823	58	1,3061	93	-1,3803
24	-0,8619	59	-0,7002	94	0,4269
25	0,2883	60	-1,6857	95	-1,0039
26	2,7951	61	-0,1464	96	0,4027
27	-2,2118	62	0,1317	97	0,4429
28	0,4876	63	0,4114	98	-0,5885
29	-0,4614	64	0,0518	99	-0,1137
30	-0,0412	65	-0,2771	100	0,2004
31	-0,8588	66	1,3764		
32	0,7909	67	1,6292		
33	-0,2455	68	-0,0258		
34	-0,8645	69	-1,5071		
35	1,8465	70	1,1779		



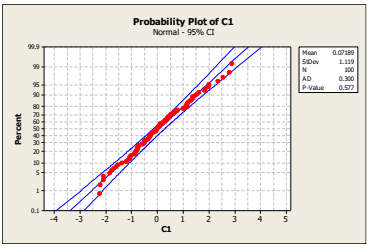
(A)



(B)



(C)



(D)

Gambar 1.2: Diagram-4 untuk data pada Tabel 1.1

Itulah Diagram-4 yang tidak boleh dilupakan pada setiap langkah awal analisis data eksploratif dan analisis statistik inferensial.

1.4. Tujuan Analisis Data Statistik

Kembali kepada analisis data eksploratif dan analisis statistik inferensial. Tujuan kedua analisis ini secara ringkas dapat dijelaskan sebagai berikut.

1.4.1. Analisis Data Eksploratif

Teknik analisis data eksploratif (EDA) tidak memerlukan Teori Probabilitas sebagai fondasinya. Yang diperlukan hanyalah logika, hermeneutika dan manipulasi aljabar (dan mungkin geometri). Namun, ia berbeda dengan teknik analisis deskriptif (*descriptive analysis* atau DA). Apabila Teknik DA hanya fokus pada pembuatan ringkasan informasi yang terkandung dalam sekelompok data, teknik EDA berupaya mengungkap informasi secara lebih luas dan dalam dengan tujuan antara lain dapat membangun hipotesis baru atau membuka area riset yang baru.

1.4.2. Analisis Statistik Inferensial

Tujuan analisis statistik inferensial (*statistical inference analysis* disingkat SIA) hanya ada 2 (dua) yakni (1) menaksir parameter, dan (2) menguji kesahihan sebuah atau beberapa hipotesis. Apabila teknik EDA tidak memerlukan Teori Probabilitas sebagai fondasinya, tidak demikian halnya dengan teknik SIA. Justru, ruh dari SIA ini adalah Teori Probabilitas yang memungkinkan para peneliti mampu mengukur kualitas hasil penaksiran parameter dan kualitas keputusan

yang diambil dalam pengujian hipotesis. Secara kongkrit, kualitas tersebut dinyatakan sebagai berikut.

1) Kualitas hasil penaksiran parameter dinyatakan sebagai "beda yang kecil (sekecil yang masih bisa ditolerir) antara nilai taksiran yang diberikan data sampel dan nilai sesungguhnya dalam populasi, dengan probabilitas (tingkat kepercayaan) yang tinggi (setinggi yang diinginkan)"
2) Kualitas keputusan pengujian hipotesis dinyatakan sebagai "kesalahan yang kecil (sekecil yang diinginkan) tatkala menolak sesuatu yang seharusnya diterima"

1.5. Organisasi

Begitu dahsyatnya efek kehadiran *outlier(s)* terhadap perilaku distribusional kelompok data yang hendak dianalisis. Bahkan, kehadiran satu buah data *outlier* saja akan berakibat fatal pada proses pengambilan keputusan. Mengingat hal tersebut, memeriksa kehadiran *outlier* dan menanganinya harus menjadi bagian rutin pada setiap analisis data dan/atau analisis statistik. Data ekstrim yang berpotensi sebagai data *outlier* harus diperiksa dengan seksama. Jika tidak ada alasan untuk percaya bahwa sebuah data ekstrim adalah outlier, maka data tersebut itu tidak boleh dipisahkan dari kelompoknya tanpa pertimbangan yang cermat. Dalam hal ini, penggunaan teknik yang tangguh (*robust*) mungkin diperlukan. Teknik yang tangguh dapat menghindari efek yang disebabkan oleh kehadiran *outlier* tanpa memisahkan dari kelompoknya.

Untuk tujuan ini, setelah Bab 1 (Pendahuluan) para pembaca akan dipandu untuk aktif terlibat dalam mendiskusikan topik-topik panas pada 12 bab berikutnya. Dimulai dengan peta fikiran dalam pembersihan data yang terdiri atas dua bab yakni Bab 2 (Langkah-langkah membersihkan data) dan Bab 3 (Outlier: Musuh dalam selimut). Kemudian dilanjutkan dengan dua bab berikutnya yang mengemukakan dua teknik penyelidikan calon tersangka outlier**;** Bab 4 (Teknik mengurutkan data) dan Bab 5 (Teknik grafikal). Selanjutnya Bab 6 (Teknik Tukey) dan Bab 7 (Iglewics-Hoaglin) adalah dua teknik penyidikan apakah calon tersangka outlier dapat dijadikan tersangka. Puncak peta fikiran berujung pada enam bab tentang enam teknik pengujian hipotesis apakah tersangka outlier

sahih sebagai outlier. Keenam bab itu adalah Bab 8 (Teknik Grubbs), Bab 9 (Teknik IESD), Bab 10 (Teknik Dixon), Bab 11 (Teknik Tietjen-Moore), Bab 12 (Teknik Roosner), dan Bab terakhir yakni Bab 13 (Teknik FMV). Harap dicatat, Bab 9 dan Bab 13 berisi dua teknik kerya kami sendiri.

Epilog dari peta fikiran, penyelidikan, penyidikan, dan pengujian hipotesis, disajikan pada bagian akhir buku ini yakni pada Bab 14 dan Bab 15. Bab 14 berisi bahan terawangan disertai dengan masalah terbuka (*open problem*) yang menantang tatkala data dipengaruhi oleh variabel waktu. Sedangkan Bab 15 berisi kumpulan data yang dapat digunakan untuk latihan.

Sebagai catatan, jika tidak ada alasan untuk percaya bahwa sebuah data ekstrim adalah *outlier*, maka data tersebut itu tidak boleh dipisahkan dari kelompoknya tanpa pertimbangan yang cermat. Dalam hal ini, penggunaan teknik yang tangguh (*robust*) mungkin diperlukan. Teknik yang tangguh dapat menghindari efek yang disebabkan oleh *outlier* tanpa memisahkan data itu dari kelompoknya.

Akhir kata, seorang analis data atau seorang statistisi tak ubahnya bagaikan sang detektif Sherlock Holmes. Analis data maupun statistisi adalah seorang pencari kebenaran saintifik, pencari fakta, pengumpul data, penyelidik, dan penyidik yang bertumpu pada kekuatan hermeneutika dan logika saintifik dengan bantuan alat-alat statistikal. Namun, ia juga sekaligus seorang jaksa dan hakim yang jujur, bijak, dan arif dalam memutuskan apakah sebuah atau beberapa data ekstrim sahih dinyatakan sebagai outlier atau tidak sahih. Dan, yang terpenting, analis data maupun statistisi adalah produsen satistik. Adapun definisi formal dari statistik adalah "variabel random yang nilainya tidak tergandung dari parameter yang tidak diketahui."

Selamat menikmati profesi sebagai detektif profesional dalam mengenalpasti outlier!

BAB 2. LANGKAH-LANGKAH MEMBERSIHKAN DATA

"*The central problem in management and leadership is failure to understand the information in variability*."

William Edward Deming

1. Keutamaan Membersihkan Data

Salah satu pilar dalam aktivitas riset saintifik berbasis statistik maupun dalam aktivitas statistik di berbagai agensi BIG (*Business, Industry, dan Government*) adalah "pembersihan data" (*data cleansing*). Pilar ini adalah pilar keempat dari tujuh pilar berikut.

1. Teori Probabilitas
   Pilar ini adalah fondasi teoritis yang bersifat sangat matematis dalam setiap pengambilan keputusan berbasis data statistik. Pilar ini menjamin bahwa setiap keputusan akan dilengkapi dengan reliabilitasnya.
2. Desain Eksperimen
   Pilar ini adalah landasan filosofis dalam merancang sampel acak disertai dengan langkah-langkah praktisnya. Perlu dicatat, sampel acak merupakan syarat untuk melaksanakan analisis data statistik. Tidak ada analisis statistik tanpa sampel acak ...!
3. Pengumpulan Data
   Proses pengumpulan data berlandaskan desain eksperimen yang handal (*reliable*) dan ditopang dengan pemahaman teori probabilitas yang kokoh akan menjamin terkumpulnya data acak yang berkualitas.
4. Pembersihan Data
   Setelah data terkumpul, sebelum dilakukan analisis, kumpulan data itu dibersihkan terlebih dahulu. Data "kotor" dipisahkan dari kelompok data lainnya. Data "kotor" tersebut bisa berupa data anomali atau data *outlier*.
5. Analisis Data Eksploratif

Sebelum terbit buku John Tukey "Exploratory Data Analysis" tahun 1977, tidak ada istilah "analisis data eksploratif" (EDA). Yang ada adalah istilah "analisis data" yang identik dengan analisis deskriptif (DA). Namun, sejak terbit buku tersebut, DA yang bersifat deskriptif ditinggalkan dan diganti oleh EDA yang bersifat eksploratif (mengeksplorasi informasi yang tersembunyi di dalam tumpukan data).

6. Analisis Statistik Inferensial
   Disebut juga "analisis konfirmatif" sebagai komplemen dari EDA. Perilaku populasi dianalisis berdasarkan perilaku sampel yang justifikasinya didasarkan kepada pilar pertama. Inferensi statistik terdiri atas dua topik yakni "teori penaksiran" (*estimation*) dan pengujian hipotesis (*hypothesis testing*). Kedua topik ini dibangun di atas landasan teori probabilitas.
7. Analisis Khusus
   Analisis khusus adalah kumpulan semua teori dan metode yang digunakan dalam aktivitas statistik. Di dalamnya ada berbagai macam analisis seperti Analisis Korespondensi, Analisis Multivariat, Analisis Non-Parametrik, Analisis Deret Waktu, Pemodelan Statistik, dlsb.

Ketujuh pilar ini tidak berdiri sendiri-sendiri, namun mereka bersifat sekuensial. Pilar kedua dibangun setelah pilar pertama. Lalu, ia diikuti oleh pilar ketiga, keempat, kelima, keenam sampai dengan pilar ketujuh. Pilar yang satu terkait erat dengan pilar yang lain.

Pilar-pilar itu akan senantiasa hadir di dalam peta fikiran (*mind-map*) setiap orang tatkala belajar statistika atau bekerja dengan menggunakan statistika atau bekerja dalam statistika (ini barang langka di Indonesia). Peta fikiran ini dimulai dari "teori probabilitas" sebagai pilar teoritis sebelum membuat desain eksperimen dan melaksanakan pengumpulan data untuk kemudian melakukan pembersihan data, EDA, analisis inferensial, dan meluncur menuju analisis khusus.

Pilar keempat yakni pembersihan data, yang identik dengan proses pemisahan data "kotor" *outlier*, adalah topik utama buku ini. Dengan

demikian, buku ini diharapkan dapat memperkokoh pemahaman para pembaca tentang peran pilar ini dalam analisis data dan analisis statistik. Untuk itu, pada bagian kedua di bawah ini, marilah kita mulai dengan topik mengenalpasti *outlier*. Namun sebelumnya, apabila pembaca berminat untuk memperkokoh pemahaman tentang pilar ketiga, silahkan baca buku Djauhari (2020) yang diterbitkan oleh ITB Press.

Dengan menggunakan peta fikiran seperti ini, aktivitas statistik menjadi atraktif, menarik, dan menyenangkan untuk digauli. Kalau pun ilmu statistika dirasakan membosankan atau sukar untuk dipelajari, itu hanya karena belum terbiasa menggunakan bahasa matematika dalam berkomunikasi dengan alam khususnya dengan data. Memang ilmu statistika manja. Bahkan amat sangat manja. Akan tetapi tidak sukar ditaklukkan. Cara menaklukkannya adalah dengan menguasai pilar pertama yang sangat matematis, dan menerapkannya pada keenam pilar lainnya yang bersifat praktis. Oleh karena itu jangan heran, semua statistisi handal adalah seorang matematisi.

2. Tiga Langkah Awal Mengenalpasti *Outlier*

Data *outlier* dan data anomali memiliki kesamaan pada nilainya yang mencurigakan. Mereka dibedakan dalam hal faktor penyebabnya. Penyebab dari kehadiran data anomali umumnya adalah kesalahan pencatatan data. Umpamanya data yang seharusnya bernilai 181 ternyata tercatat 1810. Bisa juga karena kesalahan sistem pembangkit data seperti misalnya turunnya tegangan listrik sewaktu eksperimen dilaksanakan di laboratorium sehingga menganggu kinerja alat. Atau karena sebab lain yang memungkinkan adanya koreksi terhadap data. Dalam praktik, data seperti itu dibuang saja kecuali jika ada kemungkinan untuk memperbaikinya.

Lain halnya dengan *outlier*. Sebagai data yang nilainya mencurigakan, kehadiran data *outlier* disebabkan karena kehadiran populasi yang berbeda dengan populasi yang dijadikan kajian. Umpamanya dalam survei tentang penghasilan per bulan, Seorang direktur sebuah perusahaan dengan penghasilan yang tinggi dan tinggal di suatu wilayah berpopulasi golongan menengah ke bawah tentu akan dianggap sebagai anggota populasi di luar wilayahnya

tersebut. Oleh karena itu, dia akan menjadi *outlier* di wilayah itu; data penghasilan dia harus dipisahkan dari kelompok besar data lainnya lalu dianalisis tersendiri.

Pemahaman tentang makna *outlier* secara populer dikemukakan dengan cantik di dalam Gladwell (2008). Ini buku *bestseller* yang patut dibaca oleh mereka yang ingin menjadi luar biasa (*extraordinary*) atau *outlier*. Secara statistikal, *outlier* adalah data yang kehadirannya di dalam sekelompok data akan mengganggu atau bahkan mengubah perilaku distribusional kelompok data tersebut. Dengan demikian, kehadiran *outlier* akan mengakibatkan hasil analisis data dan hasil analisis statistik yang bias. Dan, ujung-ujungnya berakibat fatal pada pengambilan keputusan yang salah dan pada penyusunan kebijakan yang keliru.

Berdasarkan uraian di atas, jelaslah, dua kemampuan berikut adalah inti dari proses pembersihan data sebelum melakukan EDA ataupun analisis statistik inferensial (SIA).

1) Memberikan label sebagai data *outlier* atau dengan kata lain menyelidiki lalu mengidentifikasi calon tersangka *outlier*. Kemudian menyidik atau mengenalpasti apakah calon tersangka dapat dinaikkan statusnya menjadi tersangka outlier. Setelah itu, menguji kesahihan apakah tersangka benar merupakan *outlier* dengan reliabiltas yang dikehendaki.
2) Mengidentifikasi data anomali kemudian memperkaikinya atau membuangnya.

Penting untuk dicatat bahwa, EDA adalah sebuah cara pandang tentang bagaimana analisis data selayaknya dilaksanakan. EDA bukan hanya menyajikan ringkasan angka-angka statistik akan tetapi jauh lebih luas dari itu. Di dalamnya termasuk bagaimana menggali informasi yang tersembunyi dalam tumpukan data, menafsirkan hasil galiannya, membangun hipotesis baru, dlsb.

Dalam praktik, proses membersihkan data terdiri atas 3 (tiga) tahap. ***Pertama*** adalah tahap penyelidikan atau identifikasi; yakni memberi label *outlier* kepada data yang dicurigai atau mengidentifikasi calon tersangka *outlier*. Proses ini biasa dilakukan dengan mengurutkan

data dan dengan visualisasi data. Data yang secara terurut atau secara visual tampak berada jauh dari kelompok besarnya, lalu diberi label calon *outlier* dan kemudian dijadikan calon tersangka *outlier*. Penguasaan teknik mengurutkan data dan teknik visualisasi data adalah tantangan utama pertama dalam pembersihan data.

Kedua adalah tahap penyidikan untuk menentukan apakah data yang dijadikan calon tersangka *outlier* layak atau tidak layak ditetapkan sebagai tersangka. Pada tahapan ini diperlukan statistik sebagai alat pengukur; artinya alat yang memberikan ukuran kuantitatif. Tidak cukup hanya mengandalkan kemampuan visual (ukuran kualitatif). Dalam praktik, seringkali alat statistik yang memadai tidak tersedia di dalam literatur. Jika demikian, kita harus buat sendiri terlebih dahulu seperti yang dilakukan oleh Willaim S. Gosset tatkala menciptakan statistik *t* yang sangat terkenal.

Kerap terjadi, alat statistik yang ada di dalam literatur sudah kuno (*obsolete*). Harus diakui, pembuatan alat statistik yang baru tidak mudah dilakukan akan tetapi bukan yang paling sukar dalam analisis data maupun analisis statistik. Para statistisi menyadari betul bahwa ini adalah tantangan utama kedua tatkala hendak menetapkan calon tersangka *outlier* menjadi tersangka.

Ketiga adalah tahap pengadilan atau pengujian hipotesis untuk memutuskan apakah tersangka secara signifikan sahih "bersalah" sebagai *outlier*. Nah, proses ini dapat dilakukan hanya apabila kita sudah berhasil menurunkan secara matematis statistik penguji yang sesuai beserta dengan distribusinya. Inilah tantangan utama ketiga dan yang paling terjal serta sukar didaki bagi para statistisi.

Catatan:
Pada tahap ketiga, hipotesis H0 (dalam ilmu statistika disebut dengan istilah hipotesis nol atau *null hypothesis*) dan alternatifnya H1 secara umum dirumuskan sebagai berikut.

H0: Tidak ada *outlier* di dalam kelompok data
H1: Ada *outlier* di dalam kelompok data tsb

Nah, untuk menguji H0 dengan alternatif H1 ini, ilmu statistika menyediakan banyak sekali statistik penguji (*statistical tests*) yang dapat digunakan. Ini menandakan bahwa proses pembersihan data atau pemisahan data *outlier* harus dilakukan dengan sangat hati-hati.

Sebagai contoh, berhati-hatilah tatkala menggunakan statistik penguji yang dibangun berdasarkan rata-rata sampel (*sample mean*) dan/atau deviasi standar sampel (*sample standard deviation*). Walaupun statistik penguji seperti ini sangat populer di kalangan para pengguna ilmu statistika, namun sangat disayangkan nilainya sangat mudah terdistorsi bahkan oleh kehadiran satu buah *outlier* sekalipun. Dengan kata lain, rata-rata sampel dan deviasi standar sampel adalah dua statistik yang tidak tangguh (*robust*).

3. Panduan Untuk Mengenalpasti *Outlier*

Tatkala membersihkan data, peta fikiran dimulai dari proses pelabelan *outlier* atau proses penyelidikan apakah data eksrim layak dijadikan calon tersangka *outlier*. Lalu diikuti dengan proses penyidikan apakah calon tersangka itu perlu ditingkatkan statusnya menjadi tersangka. Kemudian diakhiri dengan mengadili atau menguji apakah tersangka sahih dijadikan *outlier*.

Setelah memahami peta fikiran tersebut, yang menjadi pertanyaan besarnya adalah bagaimana mengimplementasikannya? Inilah topik besar yang akan kita bahas di dalam buku ini. Namun, sebelumnya perlu digarisbawahi bahwa secara umum ada empat skenario yang mungkin kita hadapi dalam setiap kegiatan membersihkan data. Keempat skenario itu adalah,

1. Skenario SS (satu-satu); satu buah *outlier* & satu variabel (*univariate*)
2. Skenario BS (banyak-satu); banyak *outlier* & satu variabel (*univariate*)
3. Skenario SB (satu-banyak); satu buah *outlier* & banyak variabel (*multivariate*)
4. Skenario BB (banyak-banyak); banyak *outlier* & banyak variabel (*multivariate*)

Pembahasan dalam buku ini akan kita batasi pada dua skenario pertama. Adapun skenario ketiga dan keempat kita simpan sebagai bahan diskusi dalam buku berikutnya tentang proses membersihkan data multivariat.

3.1. Kumpulan teknik membersihkan data

Proses membersihkan data diawali dengan proses penyelidikan atau proses mengidentifikasi calon tersangka *outlier*. Kalau ada indikasi yang kuat, maka dilanjutkan dengan menjadikannya sebagai tersangka *outlier* dan kemudian dilakukan pengujian hipotesis.

Proses penyelidikan dilaksanakan dengan:

1. Teknik berbasis data terurut; data diurutkan dari yang terkecil sampai dengan yang terbesar. Tersangka outlier akan tampak mencolok dengan nilai yang jauh berbeda dengan kelompok data lainnya. Cara ini sekaligus dapat mengidentifikasi data yang akan dijadikan tersangka.
2. Teknik grafikal; berupa visualisasi data dalam bentuk diagram kotak (boxplot), histogram, dan diagram probabilitas normal. Dapat pula ditambah dengan diagram pencar (scatter plot).

Selanjutnya, untuk menyidik apakah calon tersangka *outlier* layak dijadikan tersangka, buku ini menganjurkan penggunaan kedua teknik berikut.

1. Teknik Tukey
2. Teknik Iglewicz-Hoaglin

Setelah proses penyidikan selesai dan tersangka *outlier* telah ditentukan, maka proses berikutnya adalah mengadili atau menguji apakah tersangka sahih sebagai *outlier*. Caranya dengan melakukan pengujian hipotesis pada tingkat signifikansi yang diinginkan. Untuk itu, kita akan menggunakan enam teknik menguji yang sangat populer berikut ini.

1. Statistik penguji Grubbs (1950, 1969), yang dikenal dengan nama *Extreme Studentized Deviation* (ESD).

2. Statistik penguji IESD (*Improved Extreme Studentized Deviation*) yang diperkenalkan oleh Djauhari (2001a) dan merupakan uji ESD yang diperbaharui.
3. Statistik penguji Dixon (1950, 1960).
4. Statistik penguji Tietjen-Moore (1972).
5. Statistik penguji Rosner (1975, 1983), yang dinamakan *Generalized Extreme Studentized Deviation* (GESD)
6. Statistik penguji FMV (*Fast Minimum Variance*) yang kami kembangkan dari FMCD (*Fast Minimum Covariance Determinant*) dan dari MVV (*Minimum Vector Variance*). FMCD adalah karya Rousseeuw dan van Driessen (1999) yang sangat fundamental dan monumental. Sedangkan MVV adalah karya Herwindiati, Djauhari dan Mashuri (2007).

3.2. Karakteristik keenam statistik penguji

Secara umum, untuk skenario SS tersedia ESD dan IESD yang dapat digunakan dengan baik tatkala ukuran sampel $n \geq 7$. Lalu, untuk skenario BS ada uji Rosner yang dapat digunakan dengan baik tatkala $n \geq 20$. Ada juga uji Tietjen-Moore untuk $n \geq 30$. Selanjutnya, kami perkenalkan uji FMV untuk n mulai dari 50 sampai puluhan ribu bahkan untuk big data. Uji FMV ini kami kembangkan dari uji FMCD yang sangat terkenal dan dari uji MVV yang merupakan komplemen dari FMCD. Sebagai catatan, seperti halnya uji FMCD dan uji MVV, uji FMV adalah uji yang tangguh (*robust*).

Khusus untuk n yang kecil sampai moderat (n dari 4 s/d 25), ada uji Dixon. Namun uji ini terbatas hanya untuk menguji kehadiran 1 atau 2 buah data *outlier* saja.

Tabel 2.1. Karakteristik keenam statistik penguji

No.	Uji	Populasi	N	Skenario
1	Grubbs (ESD)	Normal	≥ 7	SS
2	Rosner	Normal	≥ 15[1]	BS
3	IESD	Normal	≥ 7[2]	SS
4	Dixon	Sembarang	4-25[3]	BS
5	Tietjen-Moore	Sembarang	≥ 7[4]	BS

6	FMV	Normal	≥ 50	BS

[1] *Menurut Rosner (1950), n ≥ 25 akurat & n ≥ 15 cukup akurat*
[2] *Seperti uji Grubbs*
[3] *Dibatasi hanya sampai 2 buah tersangka outlier*
[4] *Seperti uji Grubbs; Tietjen-Moore (1972) memberikan contoh dengan n = 8*

BAB 3. *OUTLIER*: MUSUH DALAM SELIMUT

"*It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts*."

Sir Arthur Conan Doyle, Sherlock Holmes

Proses membersihkan data tak ubahnya proses membersihkan beras tatkala hendak menanak nasi. Sebelum nasi ditanak, terlebih dahulu beras harus dibersihkan dari kehadiran kerikil-kerikil dan "kotoran-kotoran" lain. Kerikil, jelas harus kita buang. Bagaimana dengan "kotoran"? Jangan cepat-cepat buang "kotoran." Siapa tahu "kotoran" itu ternyata intan atau batu mulia lainnya. Oleh karena itu, kalau menemukan "kotoran" maka pisahkanlah. Lalu lakukan analisis tersendiri terhadapnya. Siapa tahu di dalamnya terkandung informasi yang sangat bernilai seperti tatkala Marie Currie menemukan uranium. Proses pembelajaran seperti ini merupakan inti kegiatan rutin dalam setiap upaya menjamin kebersihan dan kesehatan data statistik.

Begitu pulalah tatkala hendak melakukan EDA dan analisis statistik infetensi (SIA). Data harus dibersihkan dahulu dari berbagai "kotoran". Kalau "kotoran" itu berupa data anomali, seperti yang sering diakibatkan oleh kesalahan ketika mencatat data, maka perbaikilah (jika mungkin). Atau jika tidak mungkin bisa diperbaiki, buang saja data anomali itu. Namun, kalau "kotoran" berupa data *outlier*, maka pisahkanlah data itu dari kelompok besar data lainnya. Lalu, lakukan analisis eksploratif tersendiri terhadapnya, siapa tahu di dalamnya terkandung informasi penting dan bernilai tinggi.

Kelompok besar data yang tidak mengandung *outlier* itulah yang selanjutnya kita analisis dengan menggunakan berbagai metode statistikal yang termasuk dalam tiga pilar terakhir dari tujuh pilar aktivitas riset saintifik berbasis statistik seperti telah dikemukakan di

Bab 2. Tampak jelas bahwa proses pembersihan data (*data cleansing*) adalah sebuah keniscayaan.

Untuk memotivasi para pembaca agar senantiasa waspada terhadap kehadiran data *outlier*, di bab ini akan diberikan tiga contoh yang memperlihatkan bahaya yang akan dihadapi jika kita abai terhadap kebersihan dan kesehatan data. Harap dicatat, *outlier* adalah musuh dalam selimut; artinya ia bercampur dengan semua data lainnya dan kehadirannya tidak mudah dideteksi. Contoh pertama akan memberi pelajaran kepada kita bahwa kehadiran satu saja *outlier* bisa menurunkan kualitas analisis regresi. Yang kedua memberi peringatan bagaimana analisis deskriptif tanpa visualisasi data bisa membingungkan. Sedangkan yang ketiga memperlihatkan hasil ANOVA yang sesat hanya karena data yang dianalisis tidak dibersihkan terlebih dahulu.

1. Bahaya *Outlier* Dalam Pemodelan Regresi

Ilustrasi tentang manfaat membersihkan data *outlier* dalam analisis regresi akan diberikan melalui contoh dengan menggunakan data Forbes. Data ini, yang dapat diperoleh dari buku Weisberg (1985), terdiri atas 17 buah data acak tentang titik didih (dalam derajat Fahrenheit) dan tekanan (dalam skala inci air raksa).

Pengumpulan data itu bertujuan untuk meneliti hubungan antara X (titik didih) sebagai variabel bebas dan Y (100*Log(tekanan)) sebagai variabel tak bebas, melalui model regresi. Data disajikan pada kolom kedua (X) dan ketiga (Y) Tabel 3.1.

Pada tabel itu, kolom keempat (Y_{Pred}) adalah nilai prediksi dari Y berdasarkan persamaan regresi linear antara Y dan X, kolom kelima (e) menyatakan selisih ($Y - Y_{Pred}$) yang disebut residu, dan kolom terakhir adalah nilai statistik U untuk menguji kehadiran *outlier* dalam proses pembuatan model regresi. Statistik U ini akan digunakan untuk menguji kehadiran *outlier* berdasarkan teknik IESD yang telah disinggung di Bab 2, Sub-bagian 3.1.

Berdasarkan data pada tabel tersebut di atas diperoleh persamaan regresi $Y_{Pred} = 0{,}8955X - 42{,}131$ dengan R-Kuadrat = 99,5%. Bagi para pengguna statistik yang belum berpengalaman, nilai R-Kuadrat

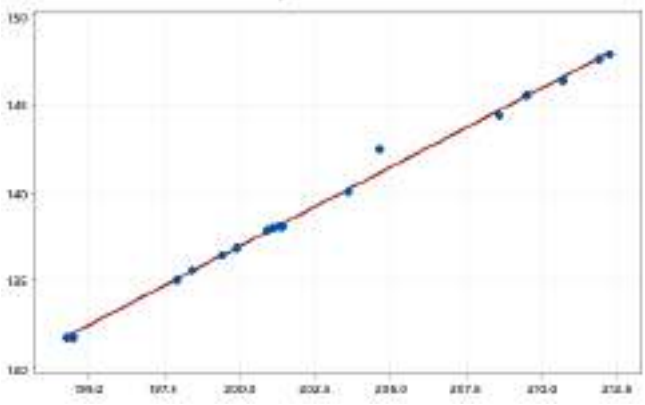
yang tinggi ini akan langsung dianggap sebagai indikasi kuat bahwa kualitas persamaan regresi itu sangat memuaskan. Apakah memang demikian? Mari kita lakukan EDA terhadap residu (e) yang ditinggalkan persamaan regresi itu.

Tabel 3.1. Data orisinal dan residu dari analisis regresi linear Y terhadap X

Observasi	X	Y	Y_{Pred}	e	U
1	194,50	131,79	132,04	-0,25	0,029926
2	194,30	131,79	131,86	-0,07	0,002226
3	197,90	135,02	135,09	-0,07	0,001834
4	198,40	135,55	135,54	0,01	0,000223
5	199,40	136,46	136,43	0,03	0,000632
6	199,90	136,83	136,88	-0,05	0,000869
7	200,90	137,82	137,77	0,05	0,001361
8	201,10	138,00	137,95	0,05	0,001408
9	201,40	138,06	138,22	-0,16	0,011889
10	201,30	138,05	138,13	-0,08	0,002826
11	203,60	140,04	140,19	-0,15	0,010421
12	204,60	142,44	141,09	1,35	0,911532
13	209,50	145,47	145,48	-0,01	0,000001
14	208,60	144,34	144,67	-0,33	0,051422
15	210,70	146,30	146,55	-0,25	0,029226
16	211,90	147,54	147,63	-0,09	0,003000
17	212,20	147,80	147,89	-0,09	0,003702

Pertama

Secara visual, hubungan $Y_{Pred} = 0,8955X - 42,131$ dapat digambarkan sebagai garis lurus pada gambar berikut.



Gambar 3.1. Garis regresi Y_{Pred} = 0,8955X – 42,131 dengan R-Kuadrat = 99,5%

Nilai R-Kuadrat = 99,5% memang sangat mempesona dan mencengangkan. Namun, mereka yang sudah berpengalaman, tidak akan cepat puas hati. Mereka selalu menginginkan hasil yang optimal. Untuk itu mereka akan dengan cermat mempertanyakan terlebih dahulu; apakah tidak ada *outlier* dalam data tersebut.

Bagaimana menjawab pertanyaan itu? Caranya adalah dengan memeriksa residu (e). Tepatnya, berdasarkan data residu itu dilakukan pengujian hipotesis H0 dengan H1 berikut.

H0: Tidak ada *outlier* dalam residu
H1: Ada *outlier* dalam residu

Dengan menggunakan uji IESD yang akan dibahas pada Bab 9, pada kolom terakhir Tabel 3.1 di atas disajikan nilai statistik penguji U. Tampak bahwa nilai U yang terkecil (0,00001) bersumber dari data nomor 13. Sedangkan yang terbesar (0,911532) sumbernya adalah data nomor 12. Adapun nilai U terbesar kedua (0,051422) berasal dari data nomor 14.

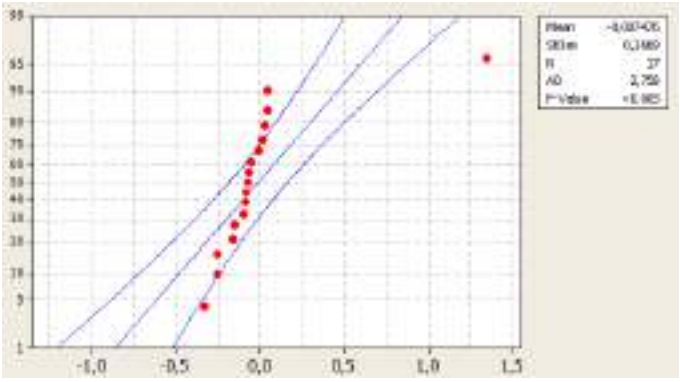
Begitu besar bedanya antara nilai U terbesar dengan nilai U terbesar kedua. Nah, ini adalah indikasi kuat bahwa data nomor 12 patut

dicurigai sebagai *outlier*. Dan, dengan menggunakan uji IESD pada tingkat signifikansi 5%, akan terbukti data tersebut adalah *outlier*. Sebagai latihan, para pembaca dipersilahkan untuk memeriksa apakah data nomor 13 yang memberikan nilai U terkecil juga berupa *outlier*.

Kedua

Temuan di atas diperkuat secara visual oleh diagram probabilitas normal untuk data orisinal yang disajikan pada Gambar 3.2 di bawah ini. Gambar ini diperoleh dengan menggunakan Minitab.

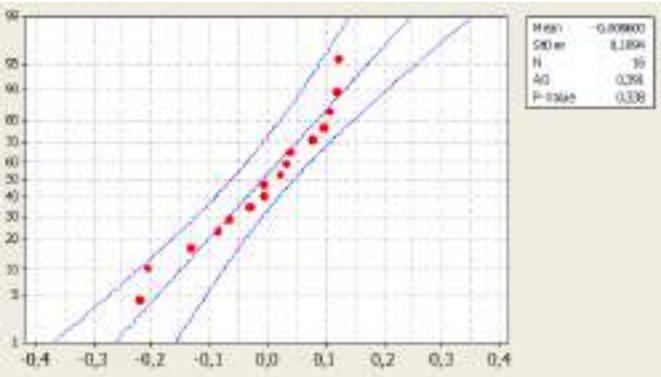
Pada diagram itu, di pojok kanan atas, tampak p-Value < 0,005 (atau 0,5%) pada pengujian kenormalan data dengan menggunakan statistik Anderson-Darling (AD). Ini berarti, untuk tingkat signifikansi 5% (yang jauh di atas nilai p-Value), residu tidak dapat dikatakan berasal dari populasi yang berdistribusi normal.



Gambar 3.2. Diagram probabilitas normal untuk residu (e) dengan daerah konfidensi 95%

Apa yang menyebabkan residu melenceng dari pola distribusi normal.? Berdasarkan analisis lebih lanjut di bawah ini, penyebabnya adalah keterlibatan data nomor 12 yang berupa *outlier* (titik berwarna merah pada Gambar 3.2 yang terisolir di sebelah kanan atas).

Mari sekarang kita analisis lebih lanjut. Apa yang akan terjadi jika data nomor 12 kita keluarkan dari kelompok data lainnya? Tanpa melibatkan data nomor 12, kita lakukan sekali lagi analisis seperti di atas untuk 16 buah data sisanya. Maka akan kita peroleh diagram probabilitas normal seperti pada Gambar 3.3.



Gambar 3.3. Diagram probabilitas normal untuk 16 residu tanpa *outlier* dengan
daerah konfidensi 95%

Gambar ini menunjukkan nilai p-Value = 0,338 (atau 33,8%). Artinya, dengan tingkat signifikansi 5% yang jauh lebih kecil dari p-Value, dengan mantap kita katakan bahwa keenambelas residu berasal dari populasi yang berdistribusi normal. Dengan kata lain, kelompok 16 data tanpa data nomor 12 sudah steril dari kehadiran *outlier* dan berdistribusi normal.

Ketiga

Selanjutnya, kita bangun lagi model persamaan regresi linear antara Y dan X berdasarkan keenam belas data yang bersih. Maka akan kita peroleh hubungan berikut.

$$Y_{Pred} = 0{,}891X - 41{,}302 \text{ dengan R-Kuadrat} = 99{,}96\%.$$

Wow … variabel tak bebas Y dan variabel bebas X memiliki hubungan linear dengan R-Kuadrat yang amat sangat tinggi! Para

pembaca pasti bukan hanya kaget, tapi mungkin juga terkesima melihat model ini yang terasa begitu amat mempesona!

Keempat
Persamaan regresi yang didasarkan kepada data bersih berbeda jauh dengan persamaan regresi yang didasarkan kepada data orisinal. Inilah gambaran tentang bahaya yang mungkin dihadapi tatkala melakukan analisis statistik berdasarkan data orisinal tanpa melalui proses pembersihan data terlebih dahulu.

2. Jebakan Anscombe
Pada tahun 1973 Francis Anscombe memperkenalkan kuartet data atau empat kumpulan data yang memiliki statistik deskriptif yang sama namun sebenarnya mereka menggambarkan empat fenomena yang berbeda.

Kuartet data itu sangat terkenal sehingga setiap orang akan dengan mudah mengunduhnya. Umpamanya, silahkan klik https://en.wikipedia.org/wiki/Anscombe%27s_quartet#. Data yang disajikan pada Tabel 3.2 diunduh dari tautan ini. Pada tabel itu tampak bahwa kuartet tersebut terdiri atas 11 buah data tentang dua variabel X dan Y yang belum tentu saling berkorelasi satu sama lain.

Tabel 3.2. Kuartet data Anscombe

Data 1		Data 2		Data 3		Data 4	
X	Y	X	Y	X	Y	X	Y
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,10	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,10	4	5,39	19	12,5

12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

Tujuan Anscombe (1973) membangun kuartet data itu adalah untuk mengingatkan kita bagaimana pentingnya visualisasi data dan pembersihan data dari kehadiran data anomali dan/atau *outlier* dalam setiap analisis statistik, Bagi para pembaca yang tertarik, berbagai diskusi mengenai kuartet data ini dapat dilihat, umpamanya, di Chatterjee dan Hadi (2006), Saville dan Wood (1991), dan Tufte (2001),

Melalui kuartet data itu, kita mendapat pelajaran tentang dua hal berikut sekaligus,

1. Bagaimana analisis deskriptif bisa mengecoh, Tepatnya, bagaimana nilai berbagai statistik seperti rata-rata, deviasi standar, koefisien korelasi, bahkan persamaan regresi bisa menipu.
2. Bagaimana visualisasi data dapat mengungkap pola data yang tidak bisa terlihat dari angka-angka statistik.

Statistik deskriptif

Seperti tampak pada Tabel 3.3, semua Data 1, Data 2, Data 3, dan Data 4 pada kuartet data di atas memiliki nilai statistik deskriptif yang sama.

Tabel 3.3. Nilai berbagai statistik keempat kumpulan data

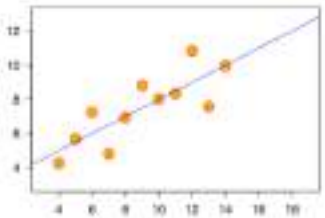
Statistik	**Nilai**
Rata-rata dari X	9
Variansi sampel dari X	11
Rata-rata dari Y	7,50
Variansi sampel dari Y	4,125
Korelasi X dan Y	0,816
Persamaan regresi	$y = 3{,}00 + 0{,}500x$
Koefisien determinasi R-Kuadrat	0,67

Pertanyaannya, apakah semua kumpulan data itu memiliki pola distribusi yang sama? Mari kita telaah lebih lanjut melalui visualisasi data.

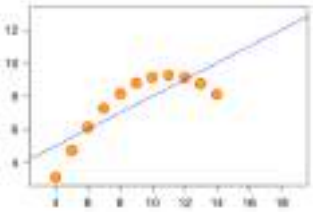
Visualisasi data

Gambar 3.4(a) – Gambar 3.4(d) bertrut-turut menampilkan diagram pencar (*scatter plot*) kuartet Data 1, Data 2, Data 3, dan Data 4. Gambar ini mempertontonkan kepada kita,

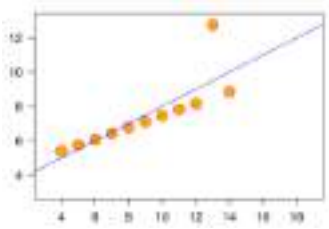
1. Sebaran data pada Gambar 3.4(a) yang mengindikasikan adanya hubungan linier antara kedua variabel X dan Y.
2. Hubungan yang tidak linear antara X dan Y pada Gambar 3.4(b). Dengan demikian, tidaklah relevan berbicara tentang koefisien korelasi Pearson. Begitu pula dengan persamaan regresi dan nilai R-Kuadrat.
3. Hubungan antara X dan Y pada Gambar 3.4(c) seyogyanya bersifat linear. Namun, kehadiran *outlier* pada data nomor 3 yakni X = 13 dengan Y = 12,74 membuat koefisien korelasi yang seharusnya bernilai 1 menjadi hanya 0,816.
4. Gambar 3.4(d) contoh data *outlier* yang dalam Analisi Regresi diberi istilah khusus yakni *leverage point*. Ini adalah data tentang variabel bebas X yang nilainya jauh berbeda dari nilai data lainnya tentang X. Dalam hal ini sebuah data X = 19 jauh berbeda dari kelompok 10 data lainnya yang berharga sama yakni X = 8. Data ini mengakibatkan adanya korelasi yang tinggi antara X dan Y (sebesar 0,816) padahal kesepuluh data lainnya menunjukkan tidak ada hubungan antara X dan Y (Everitt, 2002).



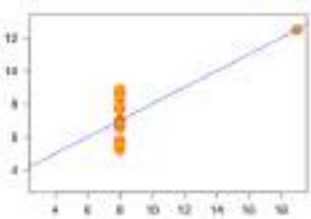
(a) Visualisasi Data 1



(b) Visualisasi Data 2



(c) Visualisasi Data 3



(d) Visualisasi Data 4

Gambar 3.4. Visualisasi kuartet data

Dari penjelasan di atas kita mendapat pelajaran bahwa kuartet data Anscombe adalah sebuah alarm agar kita berhati-hati dalam melalukan analisis statistik pada umumnya dan khususnya analisis regresi. Kalau tidak hati-hati kita bisa terjebak oleh angka-angka statistik yang mempesona. Oleh karena itu, proses pembersihan data jangan hanya mengandalkan nilai-nilai statistik saja tanpa bertumpu pada visualisasi data.

Penampilan data secara grafik sangat bermanfaat untuk mengungkap pola data, *outlier*, dan detil penting lainnya yang mungkin tidak tampak jelas dari ringkasan statistik.

3. Bahaya *Outlier* Dalam ANOVA

Pada Tabel 3.4 berikut ini disajikan data hipotetikal tentang tiga buah perlakuan P, Q, dan R. Masing-masing sebanyak n = 13, 17, dan 14 buah data acak.

Tabel 3.4. Data Orisinal

No.	Perlakuan		
	P	Q	R
1	0,0	1,2	0,0
2	0,0	1,6	0,0
3	1,2	1,7	0,0
4	1,5	2,0	1,0
5	1,6	2,0	1,0

6	2,0	2,0	1,2
7	2,2	2,1	1,2
8	2,3	2,4	1,6
9	3,7	2,6	1,8
10	4,4	2,8	1,8
11	4,7	3,4	2,3
12	4,9	3,5	2,8
13	5,5	3,8	2,9
14		4,8	5,4
15		4,9	
16		4,9	
17		5,4	

Dengan asumsi bahwa data pada setiap perlakuan berasal dari populasi berdistribusi normal, kita akan melakukan analisis variansi satu-arah (*One-Way* ANOVA) untuk menguji hipotesis H0 dengan altenatif H1 berikut.

H0: Semua perlakuan memberikan efek yang sama
H1: Ada perlakuan yang memberikan efek berbeda

Dalam ANOVA satu-arah, statistik pengujinya adalah F = A/B di mana A adalah variansi antar perlakuan (*variance between treatments*) dan B adalah variansi dalam perlakuan (*variance within treatments*). Nah, untuk memudahkan proses perhitungan F, pada Tabel 3.5 disajikan karakteristik data pada Tabel 3.4.

Tabel 3.5. Karakteristik data orisinal

Karakteristik	Perlakuan		
	P	Q	R
n	13	17	14
SUM	34,0	51,1	23,0
AVERAGE	2,61538	3,00588	1,64286
VAR	3,40474	1,78309	2,06418
Total data	44		
Grand Average	2,45682		
Variansi antar perlakuan (A)	7,36367		
Variansi dalam perlakuan (B)	2,34684		
F = A/B	3,13769		

Pada Tabel 3.5 ini kita lihat nilai statistik F sebesar F_{Sampel} = 3,13769. Sedangkan, untuk tingkat signifikansi 5%, titik kritisnya diberikan oleh tabel distribusi F dengan derajat kebebasan 2 dan 41 sebesar F_{Tabel} = 3,22568. Karena $F_{Sampel} < F_{Tabel}$, maka ANOVA menyimpulkan bahwa data sangat mendukung kesahihan hipotesis H0. Dengan kata lain, hipotesis "semua perlakuan memberikan efek yang sama" tidak dapat ditolak.

Apakah keputusan ini bijak (*wise*)? Mari kita telaah lebih seksama mengingat keputusan tersebut didasarkan kepada data orisinal yang belum mengalami proses pembersihan data terlebih dahulu. Apakah data orisinal betul-betul bersih dan sehat? Maksudnya, bersih dari kehadiran "kotoran" yang berupa *outlier* dan sehat dalam pengertian memenuhi asumsi normalitas,

Dengan menggunakan teknik-teknik pendeteksian dan pengujian *outlier* yang akan kita bahas di dalam bab-bab selanjutnya, akan kita

temukan bahwa data terbesar pada perlakuan R adalah data *outlier* untuk tingkat signifikansi 5%. Oleh karena itu, mari data ini kita sisihkan dan tidak dilibatkan dalam ANOVA lalu kita amati apa yang terjadi.

Tanpa melibatkan *outlier* tersebut, sekarang mari kita laksanakan ANOVA sekali lagi untuk data bersih yang disajikan pada Tabel 3.6. Berdasarkan data bersih ini, pada Tabel 3.7 disajikan berbagai nilai statistik termasuk statistik penguji F.

Tabel 3.6 Data bersih tanpa *outlier*

No.	Perlakuan		
	P	Q	R
1	0,0	1,2	0,0
2	0,0	1,6	0,0
3	1,2	1,7	0,0
4	1,5	2,0	1,0
5	1,6	2,0	1,0
6	2,0	2,0	1,2
7	2,2	2,1	1,2
8	2,3	2,4	1,6
9	3,7	2,6	1,8
10	4,4	2,8	1,8
11	4,7	3,4	2,3
12	4,9	3,5	2,8
13	5,5	3,8	2,9
14		4,8	
15		4,9	
16		4,9	
17		5,4	

Tabel 3.7. Karakteristik data tanpa *outlier*

Karakteristik	Perlakuan		
	P	Q	R
n	13	17	13
SUM	34,0	51,1	17,6
AVERAGE	2,61538	3,00588	1,4
VAR	3,40474	1,78309	0,96936
Total data	43		
Grand Average	2,38837		
Variansi antar perlakuan (A)	10,53277		
Variansi dalam perlakuan (B)	2,02547		
F = A/B	5,20017		

Tabel ini memberikan nilai statistik penguji F sebesar F_{Sampel} = A/B = 5,20017. Adapun titik kritisnya, untuk tingkat signifikansi 5%, diberikan oleh tabel distribusi F dengan derajat kebebasan 2 dan 40, yakni sebesar F_{Tabel} = 3,23173. Nah, setelah data *outlier* tidak dilibatkan, ternyata $F_{Sampel} > F_{Tabel}$. Ini berarti, tanpa melibatkan data *outlier*, semua data tidak mendukung kesahihan hipotesis H0. Oleh karena itu, keputusan bijak adalah menolak hipotesis yang mengatakan "semua perlakuan memberikan efek yang sama". Dengan kata lain, ada perlakuan yang memiliki efek yang berbeda.

Analisis di atas menunjukkan bagaimana berbahayanya melibatkan *outlier* dalam pengambilan keputusan berdasarkan ANOVA; bagaimana berbahayanya mengambil keputusan tanpa didahului dengan proses pembersihan data.

4, Pengujian Kenormalan Data

Pengujian kenormalan data merupakan aktivitas sentral dari setiap analisis inferensi statistikal (SIA) khususnya dalam konteks analisis statistik parametrik. Yang dimaksud dengan pengujian kenormalan

adalah pengujian hipotesis bahwa "sekelompok n buah data acak berasal dari populasi yang berdistribusi normal". Di dalam buku ini, pengujian tersebut dilakukan dengan menggunakan Minitab.

Adapun rumusan hipotesis H0 dan alternatifnya H1 secara formal adalah sebagai berikut,

H0: Data berasal dari populasi yang berdistribusi normal
H1: Data bukan berasal dari populasi yang berdistribusi normal

Lalu, bagaimana cara mengujinya? Ada banyak cara! Di antara sekian banyak cara itu, uji Anderson-Darling (AD) adalah salah satu yang paling populer. Ia digunakan secara luas dalam berbagai bidang ilmu. Popularitasnya telah mengundang minat dua peneliti dari Malaysia, Razali dan Wah (2011), untuk melakukan penelitian terhadap uji AD dan membandingkannya dengan beberapa uji yang lain. Di dalam artikel itu mereka menunjukkan keunggulan uji tersebut.

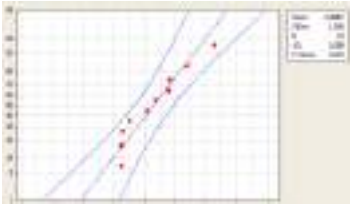
Para pembaca yang ingin mempraktikkan penggunaan uji AD, dipersilahkan menggunakan paket statistik apa saja yang banyak tersedia, Namun, untuk menunjukkan kesederhanaan proses pengujian, dalam buku ini hanya Minitab yang akan digunakan. Cara menggunakannya sebagai berikut.

1. Masuk ke dalam sistem Minitab
2. Masukkan data yang akan diuji di salah satu kolom yang tersedia. Misalnya kolom C1
3. Klik "Graph"
4. Klik "Probability Plot…"
5. Pilih "Single"
6. Lalu klik OK
7. Pada window "Graph variables:" ketiklah C1
8. Lalu klik OK
9. Sekejap kemudian, di layar monitor muncul diagram probabilitas normal
10. Selesai (tinggal membaca pesan yang diberikan diagram itu).

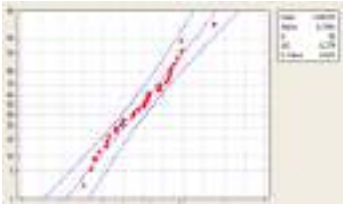
Mudah sekali, bukan? Cukup dengan membuat diagram probabilitas normal!

Contoh

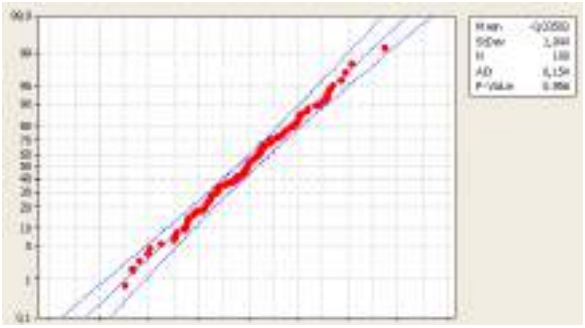
Sebagai ilustrasi, di bawah ini disajikan tiga buah gambar diagram probabilitas normal untuk data hasil simulasi dari distribusi normal standar. Gambar-gambar itu memperlihatkan data sampel dengan ukuran sampel n yang berbeda. Gambar 3.5(a) untuk n = 10 (sampel kecil), Gambar 3.5(b) untuk n = 30 (sampel moderat), dan Gambar 3.5(c) untuk n = 100 (sampel besar). Pada gambar-gambar itu, di pojok kanan atas, tertera keterangan (*legend*) yang menunjukkan nilai statistik-statistik berikut: (1) rata-rata sampel, (2) deviasi standar sampel, (3) ukuran sampel, (4) AD, dan (5) p-Value.



(a)



(b)



(c)

Gambar 3.5. Diagram probabilitas normal untuk data hasil simulasi dari distribusi normal standar dengan (a) n = 10, (b) n = 30, dan (c) n = 100

Selanjutnya, mari kita perhatikan dengan seksama gambar ini beserta keterangannya yang ada di pojok kanan atas. Maka dapat kita petik pelajaran penting berikut.

1. Baik untuk n = 10, n = 30, maupun n = 100, semua titik data berada di dalam daerah konfidensi 95%, Artinya, tidak ada indikasi kehadiran data *outlier* dalam ketiga sampel itu.
2. Pola umum data di ketiga sampel mengikuti pola garis lurus. Ini adalah indikasi bahwa hipotesis H0 untuk ketiga sampel tidak ditolak.
3. Pada semua sampel, p-Value lebih besar dari 5%. Ini bermakna kenormalan data tidak ditolak pada ketiga sampel untuk tingkat signifikansi 5%.
4. Semakin besar n, nilai AD semakin kecil dan nilai p-Value semakin besar. Tepatnya, p-Value = 0,374 untuk n = 10; 0,622 untuk n = 30; dan 0,956 untuk n = 100. Tampak bahwa, semakin banyak data hasil simulasi, perilaku distribusi data semakin mendekati distribusi populasinya dengan p-Value yang semakin mendekati 1.

Nah, informasi ini menerangkan kepada kita bahwa hipotesis "data berasal dari populasi yang berdistribusi normal" tersebut di atas tidak ditolak untuk tingkat signifikansi 5%, baik untuk sampel kecil, sampel moderat, maupun sampel besar. Hal ini tentu mudah difahami karena data pada setiap kasus sampel tersebut adalah data hasil simulasi dari distribusi normal.

5. Pelajaran Penting

Ada satu pelajaran penting dari seluruh uraian di atas yang tidak boleh diabaikan. Pelajaran itu berupa peringatan berikut. Bahaya siap menerkam para pengambil keputusan tatkala menganalisis data langsung,

1. Tanpa melalui proses pembersihan data *outlier* terlebih dahulu, dan
2. Tanpa melakukan pengujian kenormalan data (kecuali dalam analisis statistik non-parametrik).

BAB 4. MENGIDENTIFIKASI CALON TERSANGKA *OUTLIER*: TEKNIK MENGURUTKAN DATA

"*A picture is worth a thousand words*"
Confucius

Dalam upaya mengidentifikasi data yang patut dijadikan calon tersangka *outlier*, pada bab ini akan dikemukakan teknik yang paling sederhana namun banyak memberikan informasi. Teknik ini didasarkan kepada visualisasi data dengan mengurutkannya dari nilai data terkecil sampai dengan terbesar. Dengan demikian, proses identifikasi dapat dilakukan secara visual menggunakan indera penglihatan mata kepala.

Proses mengurutkan data amat sangat mudah dilakukan baik dengan menggunakan Minitab atau MS Excel. Umpamanya, di dalam MS Excel cukup digunakan fungsi SORT dengan pilihan "A to Z". Setelah diurutkan, data yang bernilai ekstrim besar (biasa disebut sebagai ekstrim kanan) dan yang bernilai ekstrim kecil (ekstrim kiri) mudah dikenali. Nah, semua data ekstrim patut dicurigai sebagai *outlier*. Namun, baru boleh dijadikan calon tersangka *outlier* hanya apabila nilainya jauh berbeda dari kelompok data lainnya. Seberapa jauh berbeda? Jawabannya akan diberikan di bab berikutnya setelah ini yakni Bab 5.

1. Keunggulan Teknik Berbasis Data Terurut

Pada Tabel 4.1 diberikan contoh data nilai matematika 18 orang mahasiswa. Kolom pertama adalah nomor urut mahasiswa dan kolom kedua menyatakan nilai mereka. Sedangkan kolom keempat memberikan daftar nilai yang sudah terurut dari yang terkecil sampai dengan yang terbesar, dan kolom ketiga adalah nomor mahasiswa yang bersesuaian dengan nilai yang sudah diurutkan.

Seketika tampak pada kolom keempat nilai data terbesar dan nilai data terkecil beserta dengan nomor mahasiswa yang bersesuaian. Inilah salah satu kelebihan teknik berbasis data terurut di mana kita dapat mengenali secara langsung nomor-nomor observasi yang

mungkin layak dijadikan calon tersangka *outlier*. Pada tabel itu tampak jelas data nomor 2 (bernilai 99) sebagai ekstrim kanan dan nomor 17 (bernilai 18) sebagai ekstrim kiri. Kedua data ekstrim ini patut dijadikan calon tersangka *outlier* mengingat nilainya yang jauh berbeda dengan nilai terdekatnya.

Tabel 4.1. Data awal dan data terurut

Data awal		Data terurut	
No.	**Nilai**	**No.**	**Nilai**
1	43	17	18
2	99	1	43
3	55	11	45
4	62	9	47
5	56	14	50
6	74	13	54
7	63	3	55
8	59	5	56
9	47	8	59
10	59	10	59
11	45	4	62
12	70	7	63
13	54	15	65
14	50	18	69
15	65	12	70
16	73	16	73
17	18	6	74
18	69	2	99

Itulah keunggulan teknik mengidentifikasi calon tersangka *outlier* dengan cara mengurutkan data. Walau demikian, patut berhati-hati pula menggunakannya mengingat berbagai kelemahannya.

2. Kelemahan Teknik Berbasis Data Terurut

Adapun kelemahan dari teknik ini adalah ketidakmampuannya dalam memberikan informasi tentang tingkat atau ukuran keberbedaan antara nilai ekstrim dengan kelompok besar data lainnya. Kelemahan ini akan kita atasi dengan menggunakan teknik Tukey dan teknik Iglewics-Hoaglin yang akan kita bahas di Bab 6 dan Bab 7.

3. Mengurutkan Data Dengan Minitab

Bagaimanakah caranya mengurutkan data dengan menggunakan Minitab? Berikut adalah instruksi dalam Minitab, yang terdiri atas 10 langkah, untuk mengurutkan data dari yang terkecil sampai dengan yang terbesar.

1. Masuk kedalam sistem Minitab
2. Simpan nomor observasi di kolom C1, dimulai dari baris ke-1
3. Simpan data di kolom C2, dimulai dari baris ke-1
4. Simpan kursor di kolom C3 baris pertama. Lalu klik ikon "Data"
5. Klik "Sort…"
6. Pada window "Column to sort by" tulis "C2" dibawah "Column"
7. Klik OK
8. Catat nomor obervasi yang sesuai dengan data terurut, pada kolom C3
9. Catat data terurut dari yang terkecil sampai dengan yang terbesar, pada kolom C4
10. Selesai.

Contoh

Untuk memperjelas kesepuluh langkah di atas, mari kita gunakan data pada Tabel 4.1.

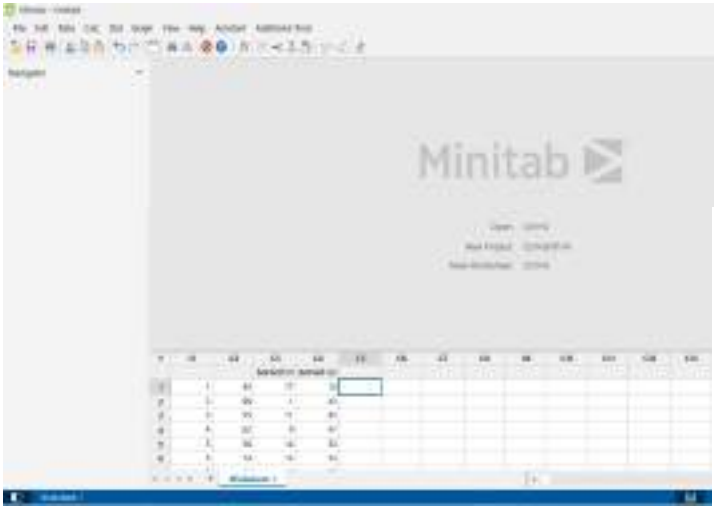
Langkah 1:	Masuk kedalam sistem Minitab (sangat jelas)
Langkah 2:	Simpan nomor observasi di kolom C1, dimulai dari baris ke-1 (COPY data pada kolom pertama Tabel 4.1 lalu
Langkah 3:	PASTE di kolom C1 pada Minitab)
Langkah 4:	Simpan data di kolom C2, dimulai dari baris ke-1 (COPY data pada kolom kedua Tabel 4.1 lalu PASTE di kolom C2 pada Minitab)
Langkah 5:	Letakkan kursor di kolom C3 baris pertama. Lalu klik ikon
Langkah 6:	"Data" (ikon ini berada di pojok kiri atas pada Gambar 4.1)
Langkah 7:	Klik "Sort…" (pilihan "Sort…" akan muncul setelah kita klik ikon "Data")

Pada jendela “Column to sort by” tulis “C2” dibawah “Column” (setelah kita klik “Sort…” akan muncul jendela (*window*) pada Gambar 4.2)
Klik OK (di bagian bawah Gambar 4.2)

Langkah 8: Catat nomor obervasi yang sesuai dengan data terurut, pada kolom C3 (sangat jelas ditampilkan pada Gambar 4.1)

Langkah 9: Catat data terurut dari yang terkecil sampai dengan yang terbesar, pada kolom C4 (sangat jelas ditampilkan pada Gambar 4.1)

Langkah 10: Selesai.



Gambar 4.1. Tampilan Minitab dan hasil pengurutan data di kolom C3 dan C4

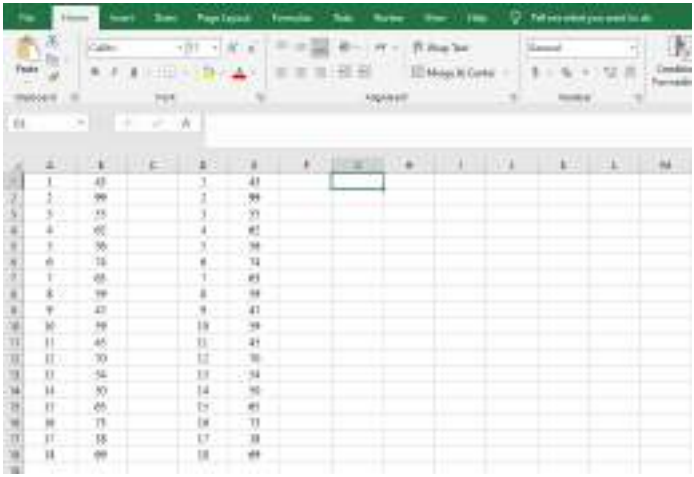
Gambar 4.2. Jendela “Column to sort by:”

4. Mengurutkan Data Dengan MS Excel

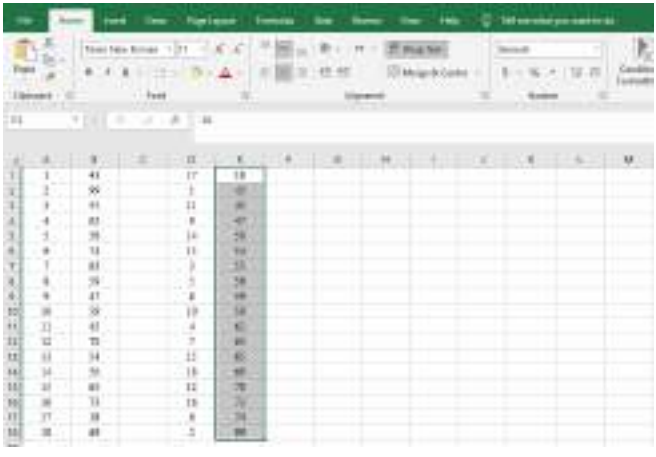
Mengurutkan data dengan menggunakan MS Excel, instruksinya sebagai berikut.

1. Masuk kedalam sistem MS Excel
2. Simpan nomor observasi, umpamanya, di kolom A1-A18 (Gambar 4.3)
3. Lalu simpan data di kolom B1-B18 (Gambar 4.3)
4. COPY data A1-A18 lalu PASTE di D1-D18 (Gambar 4.3)
5. Selanjutnya COPY data B1-B18 lalu PASTE di E1-E18 (Gambar 4.3)
6. Sorotlah (*highlight*) data E1-E18

7. Klik ikon “Sort & Filter” maka akan tampil berbagai pilihan langkah selanjutnya
8. Pilihlah “Sort Smallest to Largest” lalu klik pilihan itu
9. Pilih “Expand the Selection” lalu klik “Sort”
10. Kolom D1-D18 isinya berubah menjadi nomor obervasi yang sesuai dengan data terurut (Gambar 4.4)
11. Kolom E1-E18 isinya berubah menjadi data terurut yang yang diinginkan (Gambar 4.4)
12. Selesai.



Gambar 4.3. Tampilan layar data asli di MS Excel



Gambar 4.4. Tampilan layar data asli dan data terurut

Contoh

Dengan menerapkan keduabelas langkah di atas pada data di kedua kolom pertama Tabel 4.1, MS Excel menampilkan layar berisi data seperti tampak pada Gambar 4.3 dan menampilkan layar yang berisi data terurut pada Gambar 4.4 kolom E beserta urutan observasinya pada kolom D.

Itulah teknik pertama yang perlu diketahui untuk mengidentifikasi data yang patut dicurigai sebagai calon tersangka *outlier*. Teknik ini sangat mudah, atraktif, bahkan interaktif dan sangat membantu para peneliti setiap kali berusaha membersihkan data dari keberadaan *outlier*. Dan, dengan bantuan Minitab atau MS Excel, pekerjaan menjadi ringan dan menyenangkan.

BAB 5. MENGIDENTIFIKASI CALON TERSANGKA *OUTLIER*: TEKNIK GRAFIKAL

"*A picture is worth a thousand words but the memories are priceless*"
Fred R. Barnard

Masih dalam upaya mengidentifikasi data yang patut dicurigai sebagai calon tersangka *outlier*, pada bab ini akan dikemukakan teknik visualisasi data yang kedua selain teknik berbasis data terurut yang dibahas pada Bab 4. Basis teknik yang kedua adalah teknik grafik. Jika kedua teknik ini digunakan bersama-sama, maka proses identifikasi calon tersangka *outlier* akan memberikan hasil yang jauh lebih meyakinkan.

Ada tiga teknik utama yang patut diketahui tatkala berhadapan dengan data univariat dan satu teknik untuk data bivariat (data tentang dua variabel yang saling berkorelasi).

1. Tiga Teknik Utama

Ada tiga buah teknik utama yang tidak boleh dilupakan untuk menyelidiki keberadaan calon tersangka *outlier* di dalam sekumpulan data univariat. Ketiga teknik itu masing-masing berbasis; diagram kotak, histogram, dan diagram probabilitas (*probability plot*). Apabila ditambah dengan teknik visual yang dibahas di Bab 4, maka sudah cukuplah peralatan yang diperlukan untuk menentukan calon tersangka *outlier*. Mari kita bahas ketiga teknik tesebut.

1.1. Diagram kotak (*boxplot*)

Apa ini? Ini adalah sebuah diagram yang dibuat dengan tujuan untuk membongkar rahasia yang tersembunyi dalam sekumpulan data sampel. Caranya dengan meringkas informasi dalam bentuk gambaran tentang distribusi data.

Secara teoritis, begitu kita mengetahui bentuk distribusi dari sekumpulan data, maka tidak ada lagi rahasia yang tersembunyi dalam data itu; tidak diperlukan lagi analisis statistik. Namun dalam

praktik, seperti dikemukakan Tukey (1977), ringkasan informasi tersebut biasa ditampilkan dalam bentuk sari numerik (*numerical summary*) yang terdiri atas; data terbesar (MAX), kuartil ketiga (Q3), median (MED), kuartil pertama (Q1), dan data terkecil (MIN).

Dengan teknik ini, Minitab akan memberikan diagram kotak di mana semua data yang dicurigai sebagai calon tersangka *outlier* biasanya ditandai dengan "*asterisks*". Jadi, jika di dalam diagram ini ada tanda asterisks, itu pertanda adanya data yang patut dicurigai sebagai calon tersangka *outlier*. Jika tidak ada tanda itu, artinya Minitab tidak mampu mendeteksi. Walau demikian, amat disayangkan, diagram kotak tidak mampu menampilkan secara spesifik observasi mana yang dicurigai. Nah, disinilah teknik pengurutan data turut berperan mengidentifikasi omor observasi yang dicurigai.

Selain itu perlu dicatat bahwa, sebagaimana halnya teknik pengurutan data, diagram kotak juga tidak memberikan informasi tentang tingkat keberbedaan antara nilai data yang dicurigai dengan nilai data lainnya. Informasi tentang tingkat keberbedaan ini akan dibahas di bab yang membahas teknik Tukey.

1.1.1. Membuat diagram kotak

Untuk membuat diagram kotak sangat dianjurkan menggunakan Minitab. Berikut ini adalah 10 langkah cara membuatnya.

1. Masuk kedalam sistem Minitab
2. Simpan data di kolom C1, dimulai dari baris pertama
3. Simpan kursor di kolom C2 baris pertama
4. Klik ikon "Graph"
5. Klik "Boxplot"
6. Pilih "One Y" lalu klik "Simple"
7. Klik OK
8. Pada window "Graph variables:" tulis C1
9. Klik OK
10. Maka di layar monitor akan tampil diagram diagram kotak.

Contoh

Untuk memberikan ilustrasi bagaimana kesepuluh langkah itu membuahkan hasil, mari kita gunakan kembali data pada Tabel 4.1 di Bab 4. Pada Gambar 5.1 ditampilkan diagram kotak yang diberikan Minitab untuk data tersebut. Perhatikan pada gambar ini ada 2 tanda asterisks; yang di atas adalah MAX (data terbesar) dan yang di bawah MIN (data terkecil). Adapun batas atas kotak adalah Q3 (kuartil ketiga), dan batas bawahnya Q1 (kuartil pertama). Sedangkan segmen garis di bagian tengah kotak adalah MED (median data).

Gambar 5.1. Diagram kotak untuk data awal pada Tabel 4.1

Kedua tanda asterisks itu mengatakan bahwa kedua data ekstrim patut dicurigai sebagai calon tersangka *outlier*.

1.1.2. Kasus beberapa grup data

Kelebihan lainnya dari Minitab adalah tatkala kita berhadapan dengan beberapa grup data. Kita dapat sekaligus membuat diagram kotak untuk setiap grup data pada satu gambar. Dengan demikian, dengan mengamati gambar tersebut, kita dapat mengidentifikasi kehadiran data yang mencurigakan pada setiap grup dengan sekali pandang. Umpamanya kita berhadapan dengan 9 buah grup data. Minitab dapat membantu kita membuat 9 buah diagram kotak dalam satu gambar. Caranya sebagai berikut.

1. Masuk kedalam sistem Minitab
2. Simpan data di kolom C1, C2, …, sampai dengan C9, dimulai dari baris pertama
3. Simpan kursor di mana saja
4. Klik ikon “Graph”
5. Klik “Boxplot”
6. Pilih “Multiple Y’s” lalu klik “Simple”
7. Klik OK
8. Pada window “Graph variables:” tulis C1-C9
9. Klik OK
10. Maka di layar monitor akan tampil diagram 9 buah diagram kotak.

Contoh

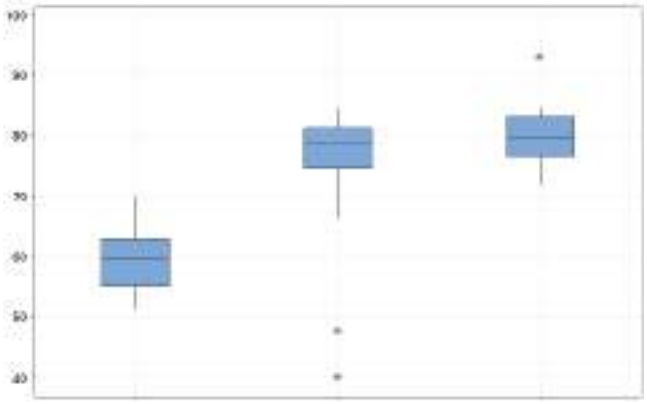
Pada Tabel 5.1 disajikan nilai matematika mahasiswa dari ketiga kelas A, B, dan C. Kelas A terdiri atas 15, kelas B ada 20, dan kelas C ada 18 orang mahasiswa. Kita hendak mengunakan teknik diagram kotak untuk mendeteksi sekaligus kehadiran calon tersangka *outlier* di ketiga kelas tersebut.

Tabel 5.1. Data nilai matematika kelas A, B, dan C

No.	Kelas		
	A	B	C
1	51	74	74
2	70	80	93
3	57	78	78
4	61	81	81
5	58	79	79
6	67	85	85
7	62	81	81
8	60	80	80
9	54	76	76
10	60	80	80
11	53	75	75

12	65	83	83
13	57	78	78
14	55	77	77
15	63	82	82
16		84	84
17		72	72
18		40	83
19		66	
20		48	

Dengan bantuan Minitab, untuk data di atas kita peroleh tiga diagram kotak berikut.



Gambar 5.2. Tiga buah diagram kotak untuk data pada Tabel 5.1

Gambar ini memperlihatkan bahwa, menurut teknik diagram kotak, di kelas A tidak ada mahasiswa yang patut dicurigai sebagai kandidat tersangka *outlier* dalam hal nilai matematikanya. Sementara itu, di kelas B ada dua mahasiswa yang patut dicurigai karena nilainya terlampau rendah. Dan, di kelas C ada satu mahasiswa calon

tersangka *outlier* dengan nilai yang amat tinggi dibandingkan dengan mahasiswa lainnya.

1.2. Histogram

Teknik histogram juga mampu menyajikan indikasi keberadaan calon tersangka *outlier* yakni dengan menampilkan batang (*bar*) frekuensi yang terisolir dari kelompok besar batang-batang frekuensi lainnya. Cara membuat histogram dengan menggunakan Minitab adalah sebagai berikut.

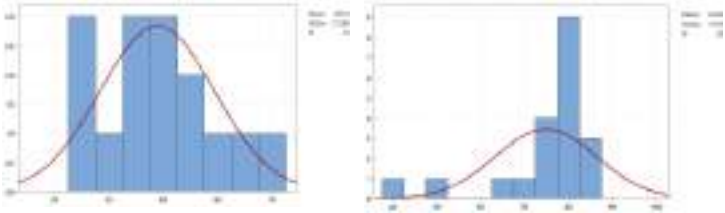
1. Masuk kedalam sistem Minitab
2. Simpan data di kolom C1, dimulai dari baris pertama
3. Simpan kursor di mana saja
4. Klik “Graph”
5. Klik “Histogram…”
6. Klik “With fit”
7. Klik OK
8. Pada window “Graph variables:” tulis C1
9. Klik OK
10. Maka di layar monitor akan tampil histogram untuk data yang ada di kolom C1

Catatan:

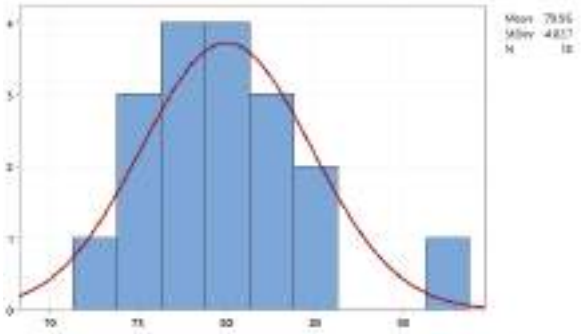
Instruksi “With fit” pada Langkah 6 dapat menampilkan histogram disertai kurva fungsi kepadatan probabilitas (*probability density function*) distribusi normal pada satu gambar. Dengan demikian kita dapat sekaligus meduga seberapa jauh distribusi data melenceng dari pola distribusi normal.

Contoh

Untuk data pada Tabel 5.1, histogram ketiga kelas A, B, dan C disajikan pada Gambar 5.3.



(a) (b)



(c)

Gambar 5.3. Hitogram (a) untuk Kelas A, (b) Kelas B, dan (c) Kelas C

Pada Gambar 5.3(a) tidak tampak adanya batang frekuensi yang terisolir. Ini menandakan, di Kelas A tidak ada data yang patut dicurigai. Tidak demikian halnya dengan Gambar 5.3 (b) dan (c). Pada Gambar 5.3(b) ada dua batang yang terisolir di sebelah kiri. Artinya, dua mahasiswa di Kelas B memiliki nilai yang rendah sekali dibandingkan dengan kawan-kawannya. Temuan ini memperkuat temuan yang diperoleh melalui teknik diagram kotak. Selanjutnya, pada Gambar 5.3(c) ada satu batang yang terisolir di sebelah kanan yang menandakan adanya mahasiswa dengan nilai ekstrim kanan dan patut dicurigai sebagai calon tersangka *outlier*. Temuan ini pun sesuai dengan hasil yang diberikan oleh diagram kotak.

1.3. Diagram Probabilitas

1.3.1. Diagram probabilitas normal

Seberapa jauh distribusi data melenceng dari pola distribusi normal? Pertanyaan ini penting untuk dijawab sebelum melakukan analisis statistik parametrik. Jawabannya dapat diperoleh dengan menggunakan diagram probabilitas normal. Keunggulan dari teknik diagram probabilitas, ia tidak hanya berguna untuk mendeteksi calon tersangka *outlier* tapi juga dapat digunakan untuk:

1. Menguji apakah data berdistribusi normal.

2. Mengukur seberapa jauh distribusi data melenceng dari distribusi normal.

Dengan menggunakan Minitab, kedua masalah ini dapat sekaligus dijawab dengan menggunakan p-Value. Untuk tingkat signifikansi 5%, kita katakan data tidak berdistribusi normal apabila p-Value < 0,05. Artinya, semakin kecil nilai p-Value dari 0,05, maka distribusi data semakin jauh dari normal. Sebaliknya, semakin besar nilai p-Value dari 0,05, distribusi data semakin dekat dengan distribusi normal.

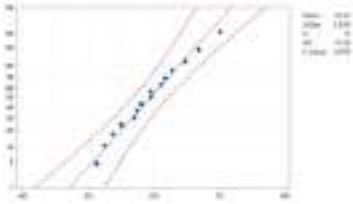
1.3.2. Membuat diagram probabilitas normal
Berikut adalah cara membuat diagram probabilitas normal dengan menggunakan Minitab.

1. Masuk kedalam sistem Minitab
2. Simpan data di kolom C1, dimulai dari baris pertama
3. Simpan kursor di mana saja
4. Klik “Graph”
5. Klik “Probability plot…”
6. Klik “Single”
7. Klik OK
8. Pada window “Graph variables:” tulis C1
9. Klik OK
10. Maka di layar monitor akan tampil diagram probabilitas normal untuk data yang ada di kolom C1.

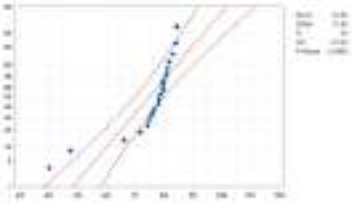
Contoh
Mari kita lanjutkan penyelidikan terhadap data nilai matematika pada Tabel 5.1 dengan mengamati diagram probabilitas normal untuk Kelas A, B, dn C. Dengan menggunakan Minitab, kita peroleh tiga diagram seperti tampak pada Gambar 5.4.

Gambar 5.4(a) memperlihatkan bahwa secara umum, titik-titik data memiliki pola garis lurus dan semuanya berada di dalam daerah konfidensi. Pola garis lurus adalah indikasi bahwa distribusi nilai di Kelas A mengikuti pola distribusi normal. Indikasi ini berubah menjadi fakta yang sangat meyakinkan setelah mengamati nilai p-

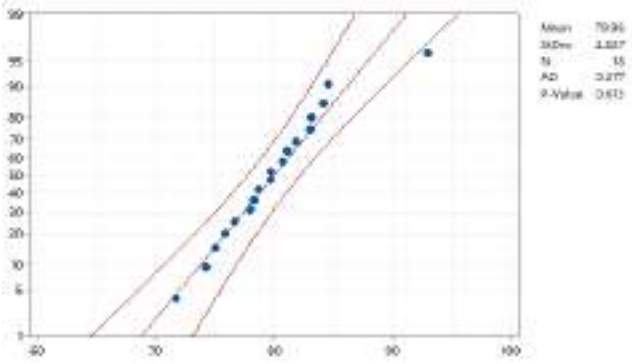
Value = 0,979 yang diberikan oleh uji Anderson-Darling (AD) seperti tampak di pojok kanan atas. Nilai ini jauh lebih besar dari 0,05.



(a)



(b)



(c)

Gambar 5.4. Diagram probabilitas normal (a) untuk Kelas A, (b) Kelas B (b), dan (c) Kelas C

Selanjutnya, kita perhatikan Gambar 5.4(b) yang menyajikan diagram probabilitas normal untuk Kelas B. Pada gambar ini titik-titik data tidak memiliki pola garis lurus. Bahkan ada beberapa data di luar daerah konfidensi. Ini pertanda, distribusi nilai di Kelas B tidak mengikuti pola distribusi normal. Bahkan, berdasarkan uji Anderson-Darling nilai p-Value jauh lebih kecil dari 0,05.

Sekarang, tibalah pada Gambar 5.4(c). Pada gambar ini titik-titik data memiliki pola garis lurus dan semuanya berada di dalam daerah konfidensi kecuali titik ekstrim kanan. Dengan p-Value = 0,613, tentu kita yakin bahwa distribusi nilai di Kelas C mengikuti pola distribusi

normal. Namun, bagaimanakah dengan data ekstrim kanan tersebut? Inilah contoh yang memperlihatkan adanya keragu-raguan dalam pengambilan keputusan berdasarkan gambar. Oleh karena itu, selain menggunakan teknik visual, baik melalui data terurut maupun melalui grafik, kita masih memerlukan adanya teknik pengujian hipotesis. Nah, teknik pengujian hipotesis adalah topik yang mendominasi buku ini dan akan kita bahas mulai Bab 8 sampai dengan Bab 13.

Sebelum kita membahas kasus bivariat, para pembaca disarankan untuk membuat diagram probabilitas normal Kelas C tanpa melibatkan data terbesar. Lalu bandingkan hasilnya dengan Gambar 5.4(c). Hasilnya sungguh mengejutkan.

2. Teknik Untuk Data Bivariat

2.1. Diagram Pencar (*Scatterplot*)

Teknik ini menyajikan data bivariat dalam bentuk diagram pada bidang dengan sumbu tegak XOY. "Bivariat" adalah kosa-kata dalam statistika yang artinya "dua variablel" yang mungkin saling berkorelasi. Artinya, pada setiap unit sampel diukur dua buah variabel X dan Y sekaligus. Nah, diagram pencar (*scatter diagram* atau *scatterplot*) mampu memperlihatkan posisi setiap titik data relatif terhadap yang lainnya pada bidang XOY.

Dengan begitu, diagram pencar dapat digunakan untuk mendeteksi keberadaan calon tersangka *outlier* pada sekelompok data bivariat. Data yang patut dicurigai sebagai calon tersangka *outlier* akan tampak secara kasat mata sebagai titik-titik data yang terisolir dari kelompok besar data lainnya.

2.2. Membuat diagram pencar

Berikut adalah instruksi dalam Minitab untuk membuat diagram pencar.

1. Masuk kedalam sistem Minitab
2. Simpan data X di kolom C1 dan data Y di kolom C2, dimulai dari baris pertama
3. Simpan kursor di mana saja
4. Klik "Graph"
5. Klik "Scatter plot…"

6. Klik “With Regression”
7. Klik OK
8. Pada window “Scatterplot With Regression” tulis C2 di kolom “Y variables” dan tulis C1 di kolom “X variables”
9. Klik OK
10. Maka di layar monitor akan tampil scatterplot yang dilengkapi garis regresi untuk data X di kolom C1 dan data Y di kolom C2.

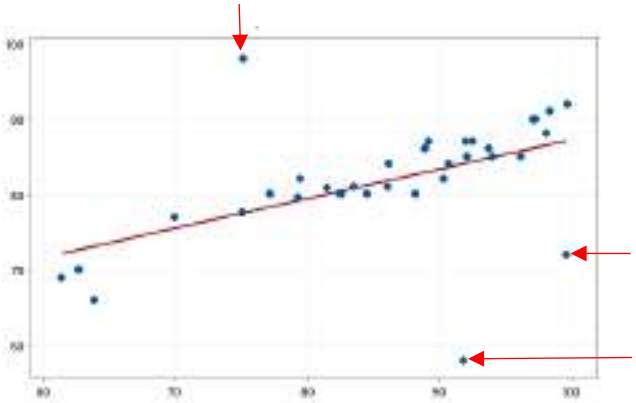
Contoh

Perhatikanlah Tabel 5.2 yang menyajikan data nilai Matematika (X) dan nilai Bahasa Indonesia (Y) dari 35 mahasiswa sebuah Perguruan Tinggi. Bagaimanakah bentuk diagram pencarnya? Lalu, adakah data yang patut dicurigai sebagai calon tersangka *outlier*?

Tabel 5.2. Data nilai Matematika (X) dan nilai Bahasa Indonesia (Y)

No.	X	Y	No.	X	Y	No.	X	Y
1	96	85	13	86	81	25	85	80
2	89	87	14	92	87	26	94	85
3	82	80	15	94	86	27	70	77
4	86	84	16	100	92	28	63	70
5	83	80	17	100	72	29	61	69
6	97	90	18	75	78	30	98	88
7	64	66	19	79	80	31	97	90
8	88	80	20	84	81	32	91	84
9	75	98	21	92	58	33	82	81
10	93	87	22	89	86	34	63	70
11	90	82	23	77	80	35	80	82
12	92	85	24	98	91			

Dengan bantuan Minitab, diagram pencar untuk data di atas disajikan pada Gambar 5.5. Gambar ini dilengkapi dengan model persamaan regresi linear $Y_{Pred} = 0{,}3908X + 48{,}15$ dan nilai R-Kuadrat = 0,2946.



Gambar 5.5. Diagram pencar untuk data pada Tabel 5.2

Mengingat sebagian terbesar data berkerumun di sekitar garis regresi, maka nilai R-Kuadrat yang kecil ini tentu mengundang pertanyaan besar. Apakah kecilnya nilai R-Kuadrat ini karena pengaruh ketiga buah data yang terisolir dari kelompoknya (ditandai panah warna merah)? Seperti telah kita bahas di dalam Bab 3, kemungkinan jawabannya adalah YA.

Kesahihannya, YA atau TIDAK, akan kita tunjukkan dengan menggunakan teknik pengujian hipotesis yang sesuai yang akan dikemukakan pada Bab 8 sampai dengan Bab 13. Walau begitu, para pembaca sangat dianjurkan untuk membuat diagram pencar beserta garis regresi dan nilai R-Kuadrat tanpa melibatkan ketiga data yang dicurigai. Setelah itu, bandingkanlah hasilnya dengan Gambar 5.5. Para pembaca pasti terkejut melihat hasilnya.

3. Catatan Penutup

Itulah tiga teknik utama yang perlu diketahui tatkala hendak menyelidiki keberadaan calon tersangka *outlier* dalam sekumpulan data uivariat, dan satu teknik untuk data bivariat. Ketiga teknik untuk data univariat itu dimulai dengan teknik diagram kotak, lalu dilanjutkan dengan teknik histogram, dan berujung pada teknik diagram probabilitas. Sedangkan untuk data bivariat diberikan teknik diagram pencar.

Pada dasarnya, keempat teknik itu bertumpu pada visualisasi data di mana calon tersangka *outlier* diidentifikasi secara kasat mata. Oleh karena keempat teknik itu tidak bertumpu pada analisis matematikal, maka penggunaannya perlu hati-hati. Dalam hal ini, penggunaan secara bersamaan sangat dianjurkan karena akan memperkaya pemahaman kita tentang keberadaan calon tersangka *outlier*.

Keempat teknik tersebut di atas adalah teknik-teknik utama yang perlu dikuasai oleh mereka yang ingin melakukan penelitian berbasis data statistik. Untuk membantu meringankan pekerjaan, dianjurkan pula menggunakan Minitab. Paket proram statistik ini murah, mudah diperoleh, sederhana dalam penggunaannya, interaktif, dan handal (*reliable*).

BAB 6. MENGENALPASTI TERSANGKA *OUTLIER*: TEKNIK TUKEY

"*Work hard to find something that fascinates you*"
Richard P. Feynman (Nobel laureate)

Ada dua teknik yang sangat populer untuk mengenalpasti apakah calon tersangka *outlier* layak atau tidak layak dijadikan tersangka. Kedua teknik ini sangat direkomendasikan penggunaannya mengingat sifatya yang tangguh (*robust*). Teknik pertama yang akan kita bahas di bab ini adalah teknik yang diperkenalkan oleh Tukey. Adapun teknik kedua, yakni teknik Iglewicz-Hoaglin, akan dikemukakan di bab berikutnya setelah yang ini.

Sejak buku "Exploratory Data Analysis" terbit tahun 1977, silakan lihat Tukey (1977) di daftar referensi, teknik Tukey amat sangat populer dan banyak digunakan di berbagai bidang. Kepopulerannya bahkan melebihi teknik Iglewicz-Hoaglin yang diperkenalkan 16 tahun kemudian di dalam Iglewicz-Hoaglin (1993). Dan, semakin tambah populer sejak Erickson dan Nosanchuk menerbitkan bukunya "Understanding Data" pada tahun 1979. Buku Erickson dan Nosanchuk (1979) ini enak dibaca dan mudah difahami bahkan oleh mahasiswa dalam rumpun ilmu-ilmu sosial khususnya dan mahasiswa di luar bidang matematika dan statistika umumnya.

1. Cara Kerja Teknik Tukey

Cara kerja teknik Tukey sangat simpel, bahkan jauh lebih sederhana ketimbang teknik Iglewicz-Hoaglin yang akan disajikan di Bab 7. Kesederhanaannya mirip dengan teknik diagram kotak. Memang, sebenarnya teknik ini merupakan pengembangan dari teknik diagram kotak. Tepatnya, teknik diagram kotak yang dilengkapi dengan dua batas pemisah antara data yang hendak dijadikan tersangka *outlier* dan kelompok data lainnya. Kedua batas itu kita sebut saja batas atas (BA) dan batas bawah (BB).

1.1. Ide Tukey

Sederhana sekali ide Tukey untuk menentukan tersangka *outlier*. Yang dia lakukan hanyalah melengkapi diagram kotak dengan BA dan BB. Namun, di balik ide sederhana itu tersembunyi dua pemikiran besar, yakni (1) bagaimana caranya menentukan tersangka *outlier*, dan (2) bagaimana menetapkan nilai BA dan BB. Inilah dua topik bahasan utama dalam bab ini.

Sebelum kita mulai membahas kedua topik tersebut, terlebih dahulu kita catat bahwa, seperti halnya teknik diagram kotak yang telah dibahas di Bab 5, teknik Tukey juga dibangun berdasarkan kelima sari numerik. Oleh karena itu, untuk sekedar pengingat, kelimanya kita tulis kembali berikut ini.

1. MIN, yakni nilai data terkecil,
2. Q1, atau kuartil pertama (25% data bernilai kurang dari atau sama dengan Q1),
3. MED, atau median (besaran yang membagi data, di kiri dan di kanan, sama banyak),
4. Q3, atau kuartil ketiga (75% data bernilai kurang dari atau sama dengan Q3), dan
5. MAX adalah data terbesar.

Nah, setelah sari numerik diperoleh, lalu kita hitung nilai ketiga statistik IQR (*interquartile range*), BA (batas atas *outlier* atau *Upper Limit*), dan BB (batas bawah *outlier* atau *Lower Limit*). Caranya sebagai berikut.

1. IQR = Q3 – Q1,
2. BA = Q3 + 1,5*IQR, dan
3. BB = Q1 – 1,5*IQR.

Berdasarkan nilai kedua statistik BA dan BB ini, calon tersangka *outlier* mungkin dapat dijadikan tersangka untuk kemudian kita tangkap.

1.2. Menangkap Tersangka *Outlier*

Setelah nilai BA dan BB sebagai ambang batas (*thresholds*) selesai dihitung, maka data yang nilainya lebih besar dari BA dan/atau lebih

kecil dari BB selanjutnya dijadikan sebagai tersangka *outlier*. Untuk lebih memahami cara kerja teknik Tukey, di bawah ini diberikan sebuah contoh.

Contoh
Mari kita gunakan lagi data tentang perlakuan P, Q, dan R pada Tabel 3.4 di Bab 3. Pada setiap kelompok data P, Q, dan R itu kita akan mencari data yang patut dijadikan tersangka dengan menggunakan teknik Tukey. Untuk itu, terlebih dahulu kita hitung sari numeriknya; MIN, Q1, MED, Q3, dan MAX. Setelah itu kita hitung IQR, BA dan BB dan kemudian kita identifkasi apakah alon tersangka layak dijadikan tersangka. Hasilnya disajikan pada Tabel 6.1 berikut.

Tabel 6.1. Sari numerik dan ambang batas outlier

Karakteristik	**Perlakuan**		
	P	**Q**	**R**
MAX	5,5	5,4	5,4
Q3	4,4	2,8	1,8
MED	2,2	2,6	1,4
Q1	1,5	2	1
MIN	0	1,2	0
IQR	2,9	0,8	0,8
BA	8,75	4	3
BB	-2,85	0,8	-0,2
Terdakwa	-	5,4*	5,4*

* *Catatan: Masing-masing ada 1 tersangka di kolom Q dan kolom R*

Baris terakhir pada tabel ini memberitahukan kepada kita bahwa:

1. Tidak ada data yang dapat dijadikan tersangka *outlier* pada perlakuan P

2. Data terbesar (5,4) pada perlakuan Q patut dijadikan tersangka
3. Demikian pula, data terbesar (5,4) pada perlakuan R layak dijadikan tersangka

2. Konstanta Pengali

Keberhasilan teknik Tukey amat tergantung dari cara menetapkan BA dan BB. Dan, bagi Tukey kedua ambang batas ini ditentukan oleh konstanta pengali sebesar 1,5. Pertanyaannya, dari mana datangnya angka 1,5 ini? Apakah angka ini mutlak? Penjelasan yang akan kami utarakan di bawah ini merupakan salah satu kontribusi dari buku ini kepada pengembangan ilmu statistika.

2.1. Menetapkan Konstanta Pengali

Berikut ini akan kami jelaskan pandangan kami tentang asal muasal munculnya angka 1,5 diikuti dengan cara menentukan nilai konstanta pengali secara umum. Namun, perlu dicatat bahwa cara yang akan kami lakukan dalam buku ini independen dengan cara yang ditempuh oleh Tukey.

Mari kita mulai dengan memperhatikan kembali kedua ambang batas BA dan BB,

$$BA = Q3 + 1{,}5*IQR \text{ dan } BB = Q1 - 1{,}5*IQR,$$

dengan IQR = Q3 – Q1. Berdasarkan formula ini Tukey menyimpulkan bahwa semua data yang nilainya lebih kecil dari BB dan/atau lebih besar dari BA layak dijadikan tersangka *outlier*.

Bagi para pembaca yang kritis, tentu yang menjadi pertanyaan adalah: "Bagaimana munculnya konstanta pengali 1,5?" Lalu, mengapa nilai konstanta pengali tersebut 1,5 dan bukan 2 atau 3, misalnya? Berikut ini kami berikan penurunan matematisnya. Mari kita mulai dengan menelusuri teori yang melatarbelakanginya.

Dalam keadaan ideal, yakni tatkala data berasal dari populasi (sebut saja X) yang berdistribusi normal dengan mean μ dan deviasi standar σ, kita memiliki hubungan probabilitas P(.) berikut,

P(X < Q1) = 0,25 dan P(X < Median) = 0,50 dan P(X < Q3) = 0,75.

Ketiga hubungan probabilitas ini berlaku umum untuk berapa pun nilai μ dan σ. Dengan demikian, ketiga hubungan itu berlaku pula untuk kasus distribusi normal standar, yakni kasus di mana $\mu = 0$ dan $\sigma = 1$. Dalam hal ini, tatkala populasinya berdistribusi normal standar (sebut saja populasi Z), maka nilai Q1, MED, dan Q3 adalah,

1. Q1 = – 0,67449 sebab P(Z < – 0,67449) = 0,25
2. MED = 0, sebab P(Z < Median) = 0,5 dan
3. Q3 = 0,67449 sebab P(Z < 0,67449) = 0,75.

Nah, dari sini kita peroleh nilai IQR = 0,67449 – (– 0,67449) = 1,34898. Akibatnya, penggunaan konstanta pengali 1,5 oleh Tukey memberikan nilai BA dan BB sebagai berikut,

BA = Q3 + 1,5*IQR = 0,67449 + 1,5*1,34898 = 2,69796, dan
BB = – BA = – 2,69796.

Di sini BB = – BA karena distribusi normal standar bersifat simetris terhadap $\mu = 0$.

Akibat lebih lanjut adalah sebagai berikut. Di bawah asumsi bahwa populasi berdistribusi normal standar, nilai BA = 2,69796 dan BB = – 2,69796 memenuhi,

P(Z < BA) = 0,996512,

dan

P(Z < BB) = 1 – P(Z < BA) = 0,003488.

Dengan kata lain, P(BB < Z < BA) = 0,996512 – 0,003488 = 0,993024.

Apa makna bilangan 0,993024 yang ada di ruas paling kanan ini? Maknanya sebagai berikut. Konstanta pengali 1,5 dipilih oleh Tukey dengan maksud agar kita hanya mentolerir kehadiran *outlier* sebanyak (100 – 99,6512)% atau sekitar 0,35% di sebelah kanan (data ekstrim kanan) dan sekitar 0,35% di sebelah kiri (data ekstrim kiri).

Jadi, secara keseluruhan tidak lebih dari 0,7% data yang merupakan *outlier*.

Itulah teori yang kami gunakan sebagai dasar pemikiran dalam memilih konstanta pengali sebesar 1,5. Dengan demikian, nilai 1,5 ini digunakan oleh Tukey karena bagi dia, cukup paling banyak 0,7% data (0,35% ekstrim kanan dan 0,35% ekstrim kiri) yang merupakan *outlier*. Oleh karena itu, konstanta pengali tersebut boleh diubah nilainya sesuai dengan tujuan penelitian. Dengan kata lain, pemilihan nilai konstanta pengali tergantung kepada berapa proporsi data yang layak dianggap sebagai *outlier*.

2.2. Kebebasan Memilih Konstanta Pengali

Apakah boleh memilih konstanta pengali yang lain selain 1,5? Tentu saja boleh! Kita bebas memilih. Namun, tergantung kepada "berapa proporsi data yang dapat kita tolerir sebagai *outlier*." Konstanta pengali itu dapat kita ubah. Bagaimana cara mengubahnya? Begini. Kita sebut saja konstanta pengali itu K. Dengan demikian, BA dan BB kita definisikan sebagai berikut,

$$BA = Q3 + K*IQR \text{ dan } BB = Q1 - K*IQR.$$

Karena IQR = Q3 – Q1, dan pada kasus distribusi normal standar berlaku Q1 = – Q3 dan BB = – BA, maka nilai BA diberikan oleh,

$$BA = (1 + 2K)*Q3.$$

Sekarang, misalkan kita tidak ingin mengikuti saran Tukey (tidak lebih dari 0,7% data yang merupakan *outlier*). Sebut saja, kita ingin 1% bukan 0,7%. Artinya, paling banyak 0,5% berupa data ekstrim kanan dan 0,5% data ekstrim kiri yang merupakan *outlier*. Maka,

$$P(Z < BA) = 1 - 0,05 = 0,995.$$

Dari persamaan ini, nilai BA dapat ditentukan dengan sangat mudah berkat bantuan MS Excel. Caranya terdiri atas lima langkah berikut,

1. Masuk ke dalam MS Excel
2. Letakkan kursor, umpamanya, di sel A1 pada Sheet 1

3. Pada sel itu tulislah (tanpa tanda " dan tanda ") "=NORM.INV(0.995,0,1)"
4. Klik ENTER
5. Di sel A1 muncul nilai BA sebesar 2,57583.

Nilai BA ini selanjutnya kita gunakan untuk menentukan nilai konstanta pengali K. Karena,

$$BA = (1 + 2K)*Q3,$$

sedangkan BA = 2,57583 dan di depan telah kita hitung Q3 = 0,67449, maka dengan demikian kita peroleh,

$$K = 1,40946.$$

Inilah konstanta pengali yang harus kita gunakan apabila tujuan penelitian menginginkan kebijakan berikut: "tidak lebih dari 1% data (0,5% data ekstrim kanan dan 0,5% data ekstrim kiri) yang merupakan *outlier*."

Nah, kebijakan penelitian itulah yang mengakibatkan nilai BA dan BB berubah menjadi,

$$BA = Q3 + 1,40946*IQR \text{ dan } BB = Q1 - 1,40946*IQR.$$

Demikianlah, statistik harus kita kembangkan tatkala kebijakan penelitian berubah. Dengan diperolehnya nilai K = 1,40946, para pembaca sekarang telah mulai belajar membuat statistik dan bukan hanya belajar statistik …!

Catatan:
Khusus bagi para pembaca yang berlatar belakang rumpun ilmu-ilmu sosial, kebijakan penelitian berikut dapat dipertimbangkan: "tidak lebih dari 10% data (5% data ekstrim kanan dan 5% data ekstrim kiri) yang merupakan *outlier*." Atau boleh juga 20% kalau mengacu kepada prinsip Pareto seperti yang dikemukakan dalam Pareto (1909) dan Newman (2006).

Nah, sebagai latihan untuk kedua kasus persentase tersebut (10% dan 20%), silahkan hitung sendiri nilai K masing-masing dengan cara seperti di atas. Lalu, setelah nilai K diperoleh, tentukanlah kedua ambang batas BA dan BB untuk setiap kasus.

Penting untuk dicatat bahwa masalah penentuan nilai K ini masih terbuka untuk dikembangkan lebih lanjut dengan memperhatikan dampak sosial dari kebijakan penelitian.

3. Komputasi BA dan BB Dengan MS Excel

Setelah nilai konstanta pengali K ditentukan, dan kelima sari numerik (MIN, Q1, MED, Q3, dan MAX) serta IQR telah dihitung, maka nilai kedua ambang batas BA dan BB akan dengan mudah dihitung dengan menggunakan MS Excel. Adapun perintah dalam sistem MS Excel terdiri atas 10 langkah di bawah ini.

Catatan:

Yang ditulis pada Langkah 3-Langkah 10, hanyalah yang tertera di antara tanda (“) dan tanda (”). Dan, konstanta pengali yang digunakan pada Langlah 8-Langkah 10 adalah K = 1,5.

1. Masuklah ke dalam sistem MS Excel
2. Misalkan kita mempunyai data acak sebanyak n = 100 yang disimpan di sel A1 sampai dengan A100
3. Di sel A101 tulislah “=QUARTILE(A1:A100,0)” lalu klik ENTER. Maka di A101 muncul nilai MIN
4. Di sel A102 tulislah “=QUARTILE(A1:A100,1)” lalu klik ENTER. Maka di A102 muncul nilai Q1
5. Di sel A103 tulislah “=QUARTILE(A1:A100,2)” lalu klik ENTER. Maka di A103 muncul nilai MED
6. Di sel A104 tulislah “=QUARTILE(A1:A100,3)” lalu klik ENTER. Maka di A104 muncul nilai Q3
7. Di sel A105 tulislah “=QUARTILE(A1:A100,4)” lalu klik ENTER. Maka di A105 muncul nilai MAX
8. Di sel A106 tulislah “=A104-A102” lalu kli ENTER. Maka di A106 muncul nilai IQR
9. Di sel A107 tulislah “=Q1-1.5*A106” lalu klik ENTER. Maka di A107 muncul nilai BB

10. Di sel A108 tulislah "=Q3+1.5*A106" lalu klik ENTER. Maka di A108 muncul nilai BA.

4. Teknik Z-Score

Mungkin pembaca pernah mendengar atau bahkan mungkin pernah menggunakan teknik Z-Score untuk menangkap tersangka *outlier*. Kami tidak terkejut jika pembaca mengatakan YA. Namun, kami betul-betul akan terkejut jika ada lembaga riset di Universitas terpandang yang menggunakan teknik Z-Score. Mengapa? Karena teknik ini tidak tangguh (*non-robust*); ia sangat sensitif terhadap kehadiran *outlier* walaupun hanya satu buah.

Dengan menggunakan data yang kami simulasi dari model campuran (*mixture model*), selanjutnya akan dipamerkan bagaimana (1) cara kerja teknik Z-Score, (2) kinerja Z-Score, dan (3) perbandingan dengan kinerja teknik Tukey.

4.1. Data Simulasi dari Model Campuran

Kinerja (*performance*) teknik Z-Score akan kita selidiki dengan cara membandingkan sensitivitasnya dengan teknik Tukey. Salah satu cara membandingkan yang biasa ditempuh para statistisi adalah dengan menggunakan data simulasi dari model campuran berikut,

$$(1 - E) *N(0,1) + E*N(\mu, \sigma^2).$$

Artinya, di dalam kelompok data itu ada (1 – E)*100% data yang berasal dari distribusi normal standar dan ada E*100% data dari distribusi normal dengan mean $\mu \neq 0$ dan/atau deviasi standar $\sigma \neq 1$ sembarang. Dengan kata lain, kita memasukkan E*100% data "kotor" ke dalam kelompok besar data sehat dan bersih dari distribusi normal standar N(0,1).

Untuk keperluan studi perbandingan kedua teknik tersebut, di sini kita gunakan nilai $\mu = 3$ dan $\sigma = 1$. Artinya, mean dari distribusi kelompok data "kotor" berada tiga deviasi standar dari mean distribusi kelompok besar data sehat dan bersih. Lalu kita bangkitkan (*generate*), umpamanya, n = 100 buah data melalui simulasi dengan E = 0,05. Dengan demikian, 95% data berasal dari N(0,1) dan 5% data

"kotor" dari N(3,1). Ini berarti ada sekitar 5 buah data yang diharapkan menjadi tersangka *outlier*. Data hasil simulasi disajikan pada Tabel 6.2.

Tabel 6.2. Data hasil simulasi dari model model campuran

No.	Data	No.	Data	No.	Data	No.	Data
1	1,13797	26	-1,94008	51	-0,80421	76	-1,16695
2	-0,25681	27	-2,17699	52	0,86779	77	1,17683
3	-0,12459	28	-0,00425	53	1,28972	78	0,52687
4	1,45161	29	-0,80865	54	0,53842	79	-0,15341
5	0,08682	30	0,15890	55	-0,45879	80	0,40181
6	-0,50389	31	0,31772	56	-2,00997	81	-1,18574
7	0,20409	32	0,92015	57	1,00463	82	0,44211
8	-0,40023	33	-0,41403	58	-0,27623	83	1,11076
9	0,59118	34	-0,81824	59	1,49955	84	-0,81696
10	0,87877	35	0,26084	60	0,23756	85	-0,18406
11	0,49851	36	-1,12203	61	0,52090	86	-0,07738
12	0,49497	37	0,72211	62	0,77951	87	-0,09965
13	-1,54310	38	0,26076	63	1,07627	88	1,76673
14	-0,07339	39	-0,88769	64	0,21230	89	-1,62630
15	-0,57143	40	0,48169	65	-0,73688	90	-0,13146
16	-0,31362	41	-0,52581	66	-0,64461	91	-0,22311
17	-1,83237	42	-1,63323	67	-0,08145	92	-1,25668
18	-1,00059	43	0,69867	68	1,06662	93	0,40197
19	-0,54240	44	-2,28350	69	-0,11204	94	-0,10403
20	-0,06710	45	1,08147	70	-0,14667	95	-1,09252
21	-0,44304	46	0,84003	71	-0,27178	96	3,26520*
22	0,33265	47	-1,44159	72	0,10757	97	3,19512*
23	-0,07446	48	0,88985	73	-0,30343	98	2,70313*
24	-0,95445	49	1,15509	74	0,27575	99	2,84559*

25	0,13609	50	-0,14474	75	-0,29264	100	2,78543*

** Catatan: 5 buah data terakhir adalah data "kotor"*

Dengan menggunakan data pada tabel ini, selanjutnya kita telaah kinerja kedua teknik itu.

4.2. Cara Kerja Teknik Z-Score

Z-Score untuk data ke-k dihitung sebagai berikut.

$$\text{Z-Score}(k) = (X_k - X_{bar})/s$$

Dalam formula ini, k bergerak dari 1, 2, …, sampai dengan n, di mana n adalah ukuran sampel. Sedangkan X_{bar} adalah rata-rata sampel, dan s deviasi standar sampel.

Teknik Z-Score umumnya digunakan di bawah asumsi "tidak lebih dari 1% data ekstrim yang merupakan *outlier* (0,5% esktrim kanan dan 0,5% ekstrim kiri)". Artinya, ambang batas atas dan ambang batas bawah adalah BA = 2,57583 dan BB = – 2,57583. Jadi, setelah Z-Score dihitung untuk semua data, selanjutnya data yang memiliki Z-Score melebihi 2,57583 dinyatakan sebagai tersangka *outlier*. Demikian pula dengan data yang memiliki Z-Score kurang dari – 2,57583.

Seperti juga teknik Tukey, ambang batas untuk Z-Score pun dapat diubah sesuai dengan kebijakan penelitian "sekian persen data yang patut dicurigai sebagai terdakwa *outlier*." Namun, dalam praktik, penggunaan teknik Z-Score bisa menyesatkan karena tidak tangguh; ia sangat sensitif terhadap kehadiran *outlier*.

Mengenai sensitivitas Z-Score, berikut ini diberikan contoh dengan menggunakan data hasil simulasi tersebut di atas.

4.3. Kinerja Teknik Z-Score

Pertama-tama kita hitung nilai Z-Score untuk setiap data hasil simulasi pada Tabel 6.2. Lalu, kita buat tabel nilai Z-Score. Hasilnya disajikan pada Tabel 6.3 berikut.

Tabel 6.3. Nilai Z-Score untuk data hasil simulasi

No.	Z-Score	No.	Z-Score	No.	Z-Score	No.	Z-Score
1	0,98400	26	-1,83985	51	-0,79778	76	-1,13056
2	-0,29559	27	-2,05720	52	0,73614	77	1,01965
3	-0,17429	28	-0,06388	53	1,12322	78	0,42337
4	1,27174	29	-0,80186	54	0,43396	79	-0,20072
5	0,01966	30	0,08579	55	-0,48089	80	0,30864
6	-0,52227	31	0,23150	56	-1,90396	81	-1,14780
7	0,12725	32	0,78417	57	0,86168	82	0,34561
8	-0,42717	33	-0,43982	58	-0,31341	83	0,95904
9	0,48237	34	-0,81065	59	1,31572	84	-0,80948
10	0,74621	35	0,17931	60	0,15796	85	-0,22884
11	0,39736	36	-1,08936	61	0,41790	86	-0,13098
12	0,39410	37	0,60248	62	0,65514	87	-0,15141
13	-1,47565	38	0,17924	63	0,92740	88	1,56084
14	-0,12732	39	-0,87437	64	0,13478	89	-1,55199
15	-0,58423	40	0,38193	65	-0,73601	90	-0,18059
16	-0,34771	41	-0,54238	66	-0,65136	91	-0,26467
17	-1,74103	42	-1,55834	67	-0,13471	92	-1,21289
18	-0,97794	43	0,58099	68	0,91855	93	0,30879
19	-0,55760	44	-2,15491	69	-0,16277	94	-0,15543
20	-0,12155	45	0,93217	70	-0,19455	95	-1,06228
21	-0,46644	46	0,71067	71	-0,30932	96	2,93556*
22	0,24519	47	-1,38253	72	0,03870	97	2,87126*
23	-0,12830	48	0,75637	73	-0,33836	98	2,41990*
24	-0,93562	49	0,99971	74	0,19299	99	2,55060*

25	0,06486	50	-0,19278	75	-0,32846	100	2,49541*

** Catatan: 5 buah data terakhir adalah data "kotor" dalam Z-Score*

Nah, sekarang kita kaji kinerja teknik Z-Score dengan mengikuti langkah-langkah berikut.

1. Urutkanlah nilai Z-Score dari yang terkecil sampai dengan yang terbesar.
2. Identifikasilah semua data yag memiliki Z-Score lebih kecil dari – 2,57583 (ekstrim kiri) dan semua data dengan Z-Score lebih besar dari 2,57583 (ekstrim kanan). Lalu, hitunglah banyaknya data ekstrim tersebut.
3. Buatlah diagram probabilitas normal untuk data nilai Z-Score sebagai bahan informasi yang bisa mendukung atau menyanggah hasil identifikasi pada Langkah 2 di atas.

Hasil pengurutan nilai Z-Score disajikan pada Tabel 6.4 (kolom OrdZ). Pada tabel ini tampak tidak ada data di ekstrim kiri yang dicurigai sebagai *outlier*. Yang ada hanyalah data di ekstrim kanan.

Tabel 6.4. Nilai Z-Score yang telah diurutkan

No.	**OrdZ**	**No.**	**OrdZ**	**No.**	**OrdZ**	**No.**	**OrdZ**
49	-2,15491	46	-0,54238	19	-0,12732	48	0,58099
32	-2,05720	11	-0,52227	25	-0,12155	42	0,60248
61	-1,90396	60	-0,48089	33	-0,06388	67	0,65514
31	-1,83985	26	-0,46644	10	0,01966	51	0,71067
22	-1,74103	38	-0,43982	77	0,03870	57	0,73614
47	-1,55834	13	-0,42717	30	0,06486	15	0,74621
94	-1,55199	21	-0,34771	35	0,08579	53	0,75637
18	-1,47565	78	-0,33836	12	0,12725	37	0,78417
52	-1,38253	80	-0,32846	69	0,13478	62	0,86168
97	-1,21289	63	-0,31341	65	0,15796	73	0,91855
86	-1,14780	76	-0,30932	43	0,17924	68	0,92740

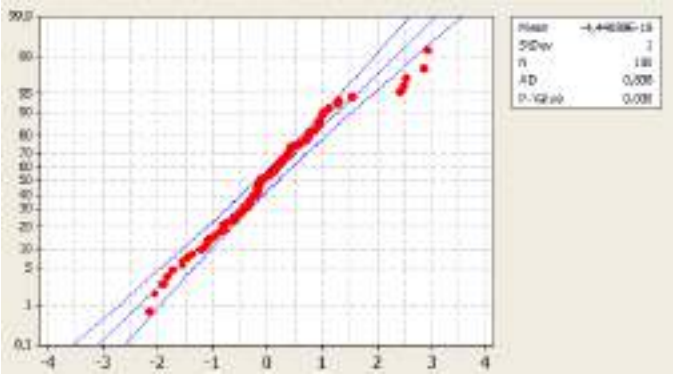
81	-1,13056	7	-0,29559	40	0,17931	50	0,93217
41	-1,08936	96	-0,26467	79	0,19299	88	0,95904
100	-1,06228	90	-0,22884	36	0,23150	6	0,98400
23	-0,97794	84	-0,20072	27	0,24519	54	0,99971
29	-0,93562	75	-0,19455	85	0,30864	82	1,01965
44	-0,87437	55	-0,19278	98	0,30879	58	1,12322
39	-0,81065	95	-0,18059	87	0,34561	9	1,27174
89	-0,80948	8	-0,17429	45	0,38193	64	1,31572
34	-0,80186	74	-0,16277	17	0,39410	93	1,56084
56	-0,79778	99	-0,15543	16	0,39736	3	2,41990
70	-0,73601	92	-0,15141	66	0,41790	5	2,49541
71	-0,65136	72	-0,13471	83	0,42337	4	2,55060
20	-0,58423	91	-0,13098	59	0,43396	2	2,87126*
24	-0,55760	28	-0,12830	14	0,48237	1	2,93556*

** Catatan: 2 data yang patut dicurigai sebagai tersangka menurut Z-Score*

Dengan menggunakan teknik Z-Score, ternyata hanya ada 2 buah data ekstrim kanan yang patut dicurigai sebagai tersangka *outlier*. Kedua tersangka ini (yakni data nomor 1 dan nomor 2) nilainya melebihi ambang batas 2,57583. Hasil yang diberikan oleh teknik Z-Score ini tentu tidak seperti yang diharapkan, karena kita tahu ada 5 buah data "kotor."

Inilah contoh yang menunjukkan bahwa teknik tersebut tidak mampu menyeret data "kotor" nomor 3, nomor 4 dan nomor 5 sebagai tersangka. Mengapa demikian? Sebagai sebuah statistik, Z-Score sangat sensitif terhadap kehadiran *outlier* sehingga nilainya mudah terdistorsi. Cukup satu saja *outlier* yang hadir, maka nilai Z-Score tidak akan terkontrol. Mengapa Z-Score sangat sensitif? Karena nilainya ditentukan oleh rata-rata sampel dan deviasi standar sampel. Kedua statistik ini terkenal sangat sensitif.

Sekarang, mari kita perhatikan dengan seksama Gambar 6.1 yang menampilkan diagram probabilitas normal untuk nilai Z-Score. Nah, seperti yang kita harapkan, pada gambar ini tampak kelima data "kotor" berada di luar daerah konfidensi. Gambar ini merupakan bantahan terhadap hasil penyidikan yang diberikan oleh teknik Z-Score.



Gambar 6.1. Diagram probabilitas normal untuk nilai Z-Score dengan tingkat konfidensi 95%

4.4. Kinerja Teknik Tukey

Sekarang mari kita bandingkan hasil penyidikan yang diberikan oleh teknik Z-Score dengan hasil dari teknik Tukey. Untuk itu data orisinal pada Tabel 6.2 kita urutkan dahulu dari yang terkecil sampai dengan yang terbesar. Hasilnya ditampilkan pada Tabel 6.5 (kolom OrdMixt).

Tabel 6.5. Data hasil simulasi yang telah diurutkan

No.	OrdMixt	No.	OrdMixt	No.	OrdMixt	No.	OrdMixt
49	-2,28350*	46	-0,52581	19	-0,07339	48	0,69867
32	-2,17699	11	-0,50389	25	-0,06710	42	0,72211
61	-2,00997	60	-0,45879	33	-0,00425	67	0,77951
31	-1,94008	26	-0,44304	10	0,08682	51	0,84003

22	-1,83237	38	-0,41403	77	0,10757	57	0,86779
47	-1,63323	13	-0,40023	30	0,13609	15	0,87877
94	-1,62630	21	-0,31362	35	0,15890	53	0,88985
18	-1,54310	78	-0,30343	12	0,20409	37	0,92015
52	-1,44159	80	-0,29264	69	0,21230	62	1,00463
97	-1,25668	63	-0,27623	65	0,23756	73	1,06662
86	-1,18574	76	-0,27178	43	0,26076	68	1,07627
81	-1,16695	7	-0,25681	40	0,26084	50	1,08147
41	-1,12203	96	-0,22311	79	0,27575	88	1,11076
100	-1,09252	90	-0,18406	36	0,31772	6	1,13797
23	-1,00059	84	-0,15341	27	0,33265	54	1,15509
29	-0,95445	75	-0,14667	85	0,40181	82	1,17683
44	-0,88769	55	-0,14474	98	0,40197	58	1,28972
39	-0,81824	95	-0,13146	87	0,44211	9	1,45161
89	-0,81696	8	-0,12459	45	0,48169	64	1,49955
34	-0,80865	74	-0,11204	17	0,49497	93	1,76673
56	-0,80421	99	-0,10403	16	0,49851	3	2,70313*
70	-0,73688	92	-0,09965	66	0,52090	5	2,78543*
71	-0,64461	72	-0,08145	83	0,52687	4	2,84559*
20	-0,57143	91	-0,07738	59	0,53842	2	3,19512*
24	-0,54240	28	-0,07446	14	0,59118	1	3,26520*

** Catatan: 5 data yang dicurigai sebagai tersangka menurut Tukey*

Untuk melihat kinerja teknik Tukey, berdasarkan data pada Tabel 6.5 kita lakukan langkah-langkah berikut.

1. Hitunglah sari numerik MIN, Q1, MED, Q3, dan MAX dari data itu. Lalu, hitung BA dan BB.
2. Periksalah apakah ada data yang bernilai kurang dari BB atau lebih dari BA. Jika ada, itulah tersangka *outlier*.

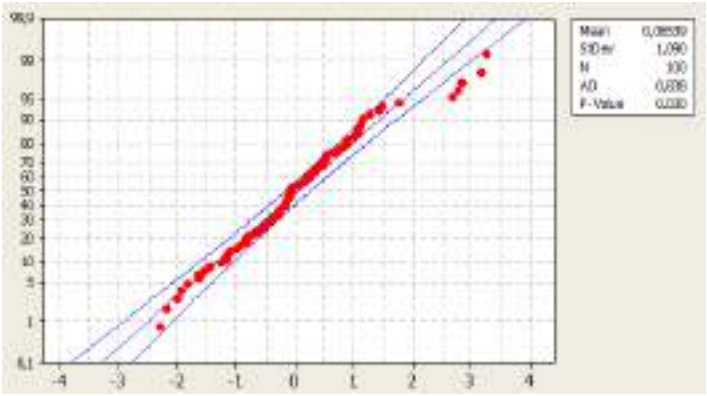
3. Buatlah diagram probabilitas normal untuk data itu untuk memperkuat atau membantah hasil yang diberikan oleh teknik Z-Score.

Dengan menggunakan teknik Tukey, diperoleh BA = 2,3401 dan BB = –2,2520. Berdasarkan teknik ini ternyata ada 6 buah data yang patut dicurigai sebagai tersangka *outlier*. Keenam terdakwa ini adalah data nomor 1 sampai dengan nomor 5 yang nilainya melebihi ambang batas atas BA = 2,3401, dan nomor 49 yang nilainya lebih kecil dari ambang batas bawah BB = –2,25198. Namun, apabila kita perhatikan dengan seksama, ekstrim kanan terdekat ke BA berjarak sekitar 0,36. Sementara itu, jarak data nomor 49 ke BB hanya sekitar 0,03.

Oleh karena itulah, hasil yang diberikan oleh teknik Tukey sangat mendekati apa yang diharapkan yakni sebanyak 5 buah data "kotor." Ini membuktikan bahwa teknik Tukey mampu menyeret kelima data "kotor" nomor 1 sampai dengan nomor 5 sebagai tersangka. Tampak jelas, teknik Tukey jauh sekali mengungguli teknik Z-Score. Namun demikian, keputusan akhir tentu akan diberikan oleh teknik pengujian hipotesis yang akan dibahas mulai Bab 8.

Adapun diagram probabilitas normal untuk data orisinal dapat kita lihat pada Gambar 6.2. Sepintas, Gambar 6.2 tampak mirip dengan Gambar 6.1. Memang demikian, karena kedua gambar ini berasal dari data yang sama di mana yang satu berasal dari data orisinal, dan yang lain berasal dari data orisinal setelah mengalami transformasi linear menjadi Z-Score. Namun, sebenarnya kedua gambar itu tidak sama. Silahkan amati nilai Mean dan StDev masing-masing di pojok kanan atas.

Nah, seperti yang kita harapkan, pada Gambar 6.2 tampak kelima data "kotor" berada di luar daerah konfidensi. Begitu pula dengan data ekstrim kiri (nomor 49) yang berada di batas bawah daerah konfidensi. Gambar ini mendukung hasil yang diberikan oleh teknik Tukey.



Gambar 6.2. Diagram probabilitas normal untuk data orisinal dengan tingkat konfidensi 95%

Itulah gambaran kinerja teknik Z-Score dibandingkan dengan teknik Tukey. Kinerja teknik Tukey jauh lebih baik ketimbang teknik Z-Score. Oleh karena itu, teknik Z-Score tidak kami rekomendasikan.

Sekarang, yang tertinggal adalah menguji kesahihan sehingga dapat diputuskan secara signifikan apakah tersangka sahih atau tidak sahih dinyatakan sebagai *outlier*. Topik kesahihan ini adalah materi yang akan mengisi lembaran-lembaran buku ini mulai dari Bab 8.

BAB 7. MENGENALPASTI TERSANGKA *OUTLIER*: TEKNIK IGLEWICZ-HOAGLIN

"Quand la statistique n'est pas fondée sur des calculs rigoureusement vrais, elle égare au lieu de diriger. L'esprit se laisse prendre aisément aux faux airs d'exactitude qu'elle conserve jusque dans ses écarts, et il se repose sans trouble sur des erreurs qu'on revêt à ses yeux des formes mathématiques de la verité."

Alexis de Tocqueville

"Manakala statistik tidak diterapkan di atas dasar perhitungan matematis yang benar-benar akurat, alih-alih membimbing, ia justru menyesatkan." – begitu kata Tocqueville. Fatwa inilah momok yang menakutkan bagi para penjahat berkedok saintifik tatkala mau memanipulasi ilmu statistik untuk kepentingannya, bukan untuk kemaslahatan umat manusia. Di sisi lain, mereka faham betul apa yang dikatakan Tocqueville selanjutnya bahwa: "Fikiran manusia mudah sekali terbuai dan terlena oleh kepalsuan akurasi statistik yang bersandar pada kesalahan terselubung bayang-bayang kebenaran matematis." Oleh karena itulah, mereka akan selalu berusaha menampilkan fondasi matematis semu dalam setiap aktivitas statistik. Namun, pada saat yang bersamaan mereka akan sekuat tenaga menghindari perdebatan matematis.

Prinsip kehati-hatian itulah yang menjiwai buku ini; kehati-hatian antara keketatan argumentasi matematis, keakuratan perhitungan dan kemaslahatan bagi umat manusia.

Tentu, dengan prinsip itu pula pada bab ini akan dikemukakan teknik kedua, yakni teknik Iglewicz-Hoaglin (1993), untuk menyelidiki apakah calon tersangka *outlier* layak atau tidak layak dijadikan tersangka. Mengingat sifatnya yang tangguh (*robust*) terhadap kehadiran *outlier*, teknik ini sangat dianjurkan penggunaannya dalam praktik. Sifat tangguh ini merupakan akibat dari cara Iglewicz-Hoaglin membangun teknik mereka yang hanya tergantung kepada median sampel. Sedangkan, dari literatur ilmu

statistika kita tahu bahwa median sampel bersifat tangguh dengan ketangguhan (*robustness*) yang maksimal.

Ketangguhannya itu harus dibayar dengan proses komputasi yang tidak sederhana. Oleh karena itu, teknik ini memerlukan kehati-hatian yang ekstra dalam penggunaannya. Ia memerlukan perhitungan yang agak jelimet. Dan, perhitungan ini harus dijamin benar sebagaimana pesan Alexis de Tocqueville tersebut di atas yang dicuplik dari Tocqueville (1909); versi bahasa Inggrisnya dapat dilihat di Lohr (1999).

Itulah salah satu tantangan dalam setiap analisis statistik; tanpa proses perhitungan yang akurat, handal, dan berkualitas tinggi, statistik dapat berubah menjadi alat ampuh untuk menipu. Walau demikian, agar supaya proses perhitungannya terasa mudah dan menyenangkan, dalam buku ini akan kita gunakan MS Excel. Paket statistik ini murah, mudah didapat, interaktif dan akrab bagi semua pemilik laptop. Dengan demikian, buku ini dapat dimanfaatkan oleh semua lapisan masyarakat dari mulai Dosen/Guru Besar, peneliti, Guru sekolah, mahasiswa, pejabat, pengusaha, polisi, jaksa, hakim, pengacara, dan Pamong serta masyarakat umum di seluruh pelosok nusantara.

1. Cara Kerja Teknik Iglewicz-Hoaglin

Cara kerja teknik ini sangat simpel karena hanya melibatkan median data. Tahun 1993, Iglewicz-Hoaglin memperkenalkan sebuah teknik menangkap tersangka *outlier* yang kami sebut teknik IH-Score. Sebenarnya teknik ini merupakan modifikasi dari teknik Z-score yang kita bahas di Bab 6 yaitu dengan mensubstitusi rata-rata sampel oleh median sampel dan deviasi standar sampel oleh kelipatan dari median deviasi absolut (*median absolute deviation*) sampel. Iglewicz-Hoaglin (1993) mendefinisikan IH-Score sebagai berikut. Untuk data ke-k, IH-Score dihitung sebagai berikut,

$$\text{IH-Score}(k) = 0{,}67449*(X_k - \text{MED})/\text{MAD}.$$

Di sini, k bergerak dari 1, 2, …, sampai dengan n, MED adalah median sampel, dan MAD adalah median deviasi absolut sampel, yakni,

MAD = Median dari ABS(X_1-MED), ABS(X_2-MED), …, ABS(X_n-MED).

Adapun ABS(X_k-MED) adalah nilai mutlak (*absolute value*) dari nilai (X_k-MED). Dengan menggunakan IH-Score sebagai detektor, Iglewicz-Hoaglin menganggap data yang memberikan nilai IH-Score melebihi 3,5 atau kurang dari – 3,5 sebagai tersangka *outlier*.

Dari mana datangnya konstanta pengali 0,67449 pada rumus IH-Score(k)? Dan, dari mana pula datangnya nilai ambang batas 3,5? Kedua pertanyaan penting tersebut akan kita bahas pada bagian akhir bab ini. Sebelumnya kita perhatikan dahulu cara kerja teknik Iglewicz-Hoaglin (disingkat teknik IH) melalui contoh berikut.

Contoh

Mari kita gunakan lagi data tentang perlakuan P, Q, dan R pada Tabel 3.4 di Bab 3. Untuk setiap kelompok data P, Q, dan R terlebih dahulu kita hitung mediannya. Diperoleh MED untuk P, Q, dan R sebesar 2,2 dan 2,6 dan 1,4. Selanjutnya pada setiap perlakuan kita hitung nilai deviasi absolut dari setiap data terhadap median sampel untuk perlakuan yang berkenaan; ABS(X_1-MED), ABS(X_2-MED), …, ABS(X_n-MED). Hasilnya disajikan pada Tabel 7.1.

Tabel 7.1. Deviasi absolut dari setiap data terhadap median sampel untuk setiap perlakuan

No.	Perlakuan		
	P	Q	R
1	2,2	1,4	1,4
2	2,2	1,0	1,4
3	1,0	0,9	1,4
4	0,7	0,6	0,4
5	0,6	0,6	0,4
6	0,2	0,6	0,2
7	0,0	0,5	0,2
8	0,1	0,2	0,2

9	1,5	0,0	0,4
10	2,2	0,2	0,4
11	2,5	0,8	0,9
12	2,7	0,9	1,4
13	3,3	1,2	1,5
14		2,2	4,0
15		2,3	
16		2,3	
17		2,8	

Nah, dari tabel ini sekarang dengan mudah kita peroleh MAD untuk P, Q, dan R sebesar 1,5 dan 0,9 dan 0,65. Dan, dengan berbekal data pada Tabel 7.1 beserta nilai MAD untuk setiap perlakuan, selanjutnya kita hitung IH-Score. Hasilnya tertera pada Tabel 7.2.

Tabel 7.2. IH-Score pada setiap perlakuan

No.	**Perlakuan**		
	P	**Q**	**R**
1	-0,98925	-1,04921	-1,45275
2	-0,98925	-0,74943	-1,45275
3	-0,44966	-0,67449	-1,45275
4	-0,31476	-0,44966	-0,41507
5	-0,26980	-0,44966	-0,41507
6	-0,08993	-0,44966	-0,20754
7	0,00000	-0,37472	-0,20754
8	0,04497	-0,14989	0,20754
9	0,67449	0,00000	0,41507
10	0,98925	0,14989	0,41507
11	1,12415	0,59955	0,93391

12	1,21408	0,67449	1,45275
13	1,48388	0,89932	1,55652
14		1,64875	4,15071*
15		1,72370	
16		1,72370	
17		2,09841	

Pada tabel itu tampak bahwa tidak ada data yang patut dijadikan tersangka pada perlakuan P dan Q. Tidak demikian halnya dengan perlakuan R; IH-Score dari data terbesar pada perlakuan R adalah 4,15071 yang lebih besar dari 3,5. Jadi, data terbesar pada perlakuan R (yakni sebesar 5,4) patut dijadikan tersangka *outlier* berdasarkan teknik IH-Score.

2. Konstanta Pengali

2.1. Menentukan Nilai Konstanta Pengali

Dari mana datangnya konstanta pengali sebesar 0,67449 pada formula untuk menghitung IH-Score(k)? Mari kita kaji lebih lanjut. Untuk diketahui pembaca, kajian yang kami lakukan ini independen dengan kajian yang dilakukan Iglewicz-Hoaglin.

Harap dicatat bahwa, manakala data berasal dari populasi berdistribusi normal, MED = μ. Sedangkan deviasi standar σ oleh Iglewicz-Hoaglin didekati menggunakan MAD/Q3. Dengan demikian, IH-Score tidak lain adalah Z-score yang dimodifikasi sehingga menjadi *robust*. Inilah alasan mengapat IH-Score sering disebut *modified Z-score*.

Telah dikemukakan di Bab 6, Bagian 2.1, bahwa dalam keadaan ideal di mana kita berhadapan dengan distribusi normal standar, maka Q3 = 0,67449. Sekarang tampak jelas bahwa konstanta pengali 0,67449 pada IH-Score tidak lain adalah Q3 jika data berasal dari distribusi normal.

2.2. Ambang Batas

Nilai ambang batas yang digunakan untuk menangkap tersangka *outlier* dengan menggunakan teknik IH ditetapkan 3,5. Tentu pembaca bertanya-tanya; dari mana datangnya nilai ambang batas ini? Di Bab 6 telah kita pelajari bahwa konstanta pengali yang digunakan pada teknik Tukey berkaitan dengan kebijakan penelitian "dalam sekumpulan data ada tidak lebih dari 0,7% data yang merupakan *outlier*."

Nah, sekarang kita selidiki kebijakan seperti apa yang diterapkan oleh Iglewicz-Hoaglin sehingga mereka mendefinisikan ambang batas sebesar 3,5. Mari kita mulai dengan mengasumsikan bahwa data berasal dari distribusi normal standar. Dalam keadaan ideal ini, maka

$$P(Z < 3,5) = 0,999767.$$

Ini mengakibatkan $1 - P(Z < 3,5) = 0,000233$ dan $P(-3,5 < Z < 3,5) = 0,999767 - 0,000233 = 0,999535$. Jadi, proporsi populasi yang ada di sebelah kiri – 3,5 dan di sebelah kanan 3,5 adalah $(1 - 0,999535)*100\%$ atau $0,000535*100\%$ atau sekitar 0,05%. Proporsi ini dapat diartikan sebagai berikut: "dalam sekumpulan data ada tidak lebih dari 0,05% data yang merupakan *outlier*."

Kalau ingin dibuat kebijakan penelitian yang sama seperti dalam teknik Tukey yakni "dalam sekumpulan data ada tidak lebih dari 0,7% data yang merupakan *outlier*," maka ambang batas atas pada Teknik Iglewicz-Hoaglin bukan 3,5 tapi BA yang memenuhi (lihat nilai BA pada teknik Tukey di Bab 6),

$$P(Z < BA) = 1 - 0,7\%/2 = 1 - 0,0035 = 0,9965$$

yang memberikan BA = 2,69796 atau dibulatkan menjadi 2,7 dan BB = – BA = – 2,7. Oleh karena itu, untuk kebijakan penelitian yang sama dengan kebijakan pada Teknik Tukey, ambang batas atas dan ambang batas bawah pada teknik IH-Score adalah 2,7 dan – 2,7 bukan 3,5 dan – 3,5.

3. Komputasi IH-Score

Ada dua statistik yang harus dihitung terlebih dahulu nilainya sebelum menentukan IH-Score untuk setiap data dalam sekelompok data yang hendak dibersihkan. Kedua statistik itu adalah median sampel MED dan median deviasi absolut sampel MAD.

Dengan bantuan MS Excel, cara menghitung MED adalah sebagai berikut.

1. Masuklah ke dalam sistem MS Excel
2. Simpan n buah data yang hendak dibersihkan dalam salah satu kolom, misalnya kolom A. Kalau n = 100, simpanlah data itu umpamanya di A1 sampai dengan A100.
3. Letakkan kursor di salah satu sel/kotak kosong, misalnya A101. Lalu, di sel tersebut tulislah perintah berikut (tanpa tanda " dan tanda "): "=MEDIAN(A1:A100)"
4. Maka di sel A101 akan tampak nilai MED dari semua data yang disimpan di A1 – A100.

Selanjutnya, kita hitung nilai MAD dan nilai IH-Score. Untuk itu, ikutilah Langkah 5 sampai dengan Langkah 9 berikut.

5. Letakkan kursor di B1. Lalu, tulislah perintah berikut (yang ada di antara tanda " dan tanda "): "=ABS(A1-$A$101)"
6. Letakkan kursor di pojok kanan bawah sel B1 sampai muncul tanda + berwarna hitam. Lalu, klik dua kali berturut-turut serentak. Maka di kolom B1 – B100 akan muncul 100 buah data tentang deviasi ablsolut
7. Letakkan kursor di B101. Lalu, tulislah perintah berikut (tanpa tanda " dan tanda "): "=MEDIAN(B1:B100)" Maka di sel B101 akan tampil nilai MAD.
8. Letakkan kursor di C1. Lalu tulislah (tanpa tanda " dan tanda ") "=0,67449*B1/$B$101"
9. Letakkan kursor di pojok kanan bawah sel C1 sampai muncul tanda + berwarna hitam. Lalu, klik dua kali berturut-turut serentak. Maka di kolom C1 – C100 akan muncul 100 buah data tentang IH-Score dari keseratus data yang hendak dibersihkan.

Sederhana, bukan? Selamat mencoba!

4. Peringatan Tentang Teknik Z-Score

Dalam praktik, masih banyak dijumpai orang yang menggunakan teknik Z-score ketimbang teknik Tukey atau teknik IH-Score. Padahal, teknik Z-score mudah menyesatkan. Mengapa? Selain teknik Z-Score tidak tangguh, secara umum telah dibuktikan oleh Iglewicz-Hoaglin (1993) bahwa nilai maksimum Z-score tidak akan melebihi (n−1)/SQRT(n). Jadi, umpamanya, manakala n = 9 maka nilai Z-score tidak akan pernah melebihi +/- 2,7. Tentu saja perolehan ini tidak masuk akal (*absurd*), bukan? Mengapa *absurd*? Karena, dalam praktik, nilai Z-score bisa lebih dari 3 untuk n berapa saja.

Catatan**:**
SQRT(n) adalah kalimat MS Excel yang menyatakan akar pangkat dua dari n.

BAB 8. UJI KESAHIHAN: TEKNIK GRUBBS

"*All models are wrong but some are useful*."

George E.P. Box

Para profesional di bidang kualitas (*quality professionals*) dari latar belakang bidang ilmu apapun selalu menyandarkan diri pada data yang sehat dan bersih dari kehadiran data *outlier*; artinya pada data yang siap dianalisis. Nah, untuk mendapatkan data yang bersih, maka pemilihan statistik penguji yang berkualitas dan penggunaannya yang tepat merupakan langkah sangat fundamental. Kalau untuk itu mereka memilih statistik penguji Grubbs (1950), yang lebih dikenal dengan nama *Extreme Studentized Deviation* (disingkat ESD), mereka berada pada jalan yang benar. Mengapa? Karena,

1. ESD direkomendasikan oleh American Society for Quality (lihat Iglewicz dan Hoaglin (1993)).
2. ESD memiliki kuasa (*power*) yang optimal tatkala menguji satu (*single*) *outlier* seperti dikemukakan dalam Tietjen dan Moore (1972) dan Rosner (1975, 1983).
3. ESD mudah dihitung. Proses komputasinya jauh lebih sederhana ketimbang statistik penguji *outlier* lainnya (Tietjen dan Moore (1972, 1973)).
4. ESD sangat populer dan digunakan secara luas. Bahkan baru-baru ini beberapa peneliti berupaya meningkatkan kinerja Uji Grubbs (Adikaram, et al. (2015) dan Djauhari (2001a)).

1. Cara Kerja Teknik Grubbs

Bayangkahlah kita berhadapan dengan data acak sebanyak n yang hendak kita bersihkan dari kehadiran *outlier*. Kita sebut saja data acak itu $X_1, X_2, \ldots, X_n$. Dan, kita asumsikan data itu diambil dari populasi yang berdistribusi normal. Nilai uji Grubbs kita hitung dengan mengikuti langkah-langkah berikut.

1. Hitung rata2 data (X_{bar}).
2. Hitung standar deviasi data (s).

3. Hitung $W_1 = ABS((X_1 - X_{bar})/s)$, $W_2 = ABS((X_2 - X_{bar})/s)$, …, $W_n = ABS((X_n - X_{bar})/s)$. Sekali lagi, ABS(.) adalah harga mutlak dari nilai yang ada di dalam tanda kurung.
4. Urutkan nilai W_1, W_2, …, W_n dari yang terkecil sampai dengan terbesar. Hasil pengurutan ini kita sebut saja $W_{(1)}$, $W_{(2)}$, …, $W_{(n)}$.
5. Maka, nilai statistik penguji ESD adalah $W_{(n)}$ yakni nilai terbesar di antara W_1, W_2, …, W_n.

Di depan telah dikemukakan bahwa ESD memiliki kuasa (*power of the test*) yang optimal tatkala menguji satu (*single*) *outlier*. Walaupun demikian, dalam praktik ESD dapat digunakan pula untuk menguji kehadiran beberapa data tersangka *outlier* asalkan tidak terjadi *masking effect* dan/atau *swamping effect*. Caranya dengan menguji *outlier* satu-per-satu dimulai dengan data yang paling ekstrim. Dan, sebagai hal khusus dari teknik Tietjen-Moore (akan dibahas pada Bab 11), ESD dapat digunakan untuk $n \geq 7$. Secara spesifik, Tietjen-Moore (1972) memberikan contoh dengan $n = 8$.

Adapun yang dimaksud dengan *masking effect* adalah keadaan di mana data *outlier* yang satu "melindungi" yang lain sehingga semua data *outlier* yang hadir tidak dapat dideteksi. Sedangkan *swamping effect* adalah keadaan di mana data yang bukan *outlier* dinyatakan sebagai *outlier*. Contoh berikut akan memberikan ilustrasi bagaimana ESD bekerja.

Contoh:
Data pada kolom kedua di Tabel 8.1 adalah tentang nilai kolesterol 15 orang yang sehat. Data tersebut dipinjam dari Bolton (1990) halaman 356. Dengan pengetahuan bahwa data berasal dari populasi yang distribusinya mendekati distribusi normal, Bolton menguji apakah kedua data ekstrim 165 dan 297 merupakan *outlier*. Untuk itu, Bolton menggunakan statistik penguji Dixon (1950).

Tabel 8.1. Data dan nilai statistik W

No.	Data	W
1	165	1,63548
2	188	0,89110

3	194	0,69691
4	197	0,59982
5	200	0,50273
6	202	0,43710
7	205	0,34090
8	210	0,17908
9	214	0,04963
10	215	0,01726
11	227	0,37111
12	231	0,50057
13	239	0,75948
14	249	1,08313
15	297	2,63662

Sementara itu, Djauhari (2001a, 2001b, 2003) tidak menggunakan Dixon tapi ESD. Mengapa? Sebab, sesuai laporan dalam Iglewicz dan Hoaglin (1993), statistik penguji Dixon hanya untuk ukuran sampel n yang sangat kecil. Berikut adalah hasil penyelidikan, penyidikan, dan pengujian.

Hasil riset yang dilakukan Djauhari (2001a, 2001b) terhadap data itu, dengan menggunakan teknik Tukey, menunjukkan:

1. Data ekstrim 165 dan 297 patut dicurigai sebagai 2 calon tersangka *outlier* mengingat letaknya yang jauh dari data terdekatnya masing-masing.
2. Data 297 jauh melebihi batas atas BA (255,75) sedangkan data 165 hanya sedikit lebih besar dari batas bawah BB (164,25). Jadi, 297 layak dijadikan tersangka seperti tampak pada Gambar 8.1 (tanda asterisks *). Sementara 165 cukup meragukan untuk dijadikan terdakwa mengingat letaknya hanya sedikit di atas ambang batas bawah.
3. Tabel 8.1 kolom ketiga yang berisi hasil perhitungan nilai statistik W_1, W_2, …, W_{15}.

4. Tabel 8.2 kolom terakhir yang memperlihatkan $W_{(1)}$, $W_{(2)}$, …, $W_{(15)}$ yakni hasil pengurutan kelima belas nilai pada Tabel 8.1 kolom ketiga, diurutkan dari yang terkecil sampai dengan yang terbesar. Pada Tabel 8.2 tampak bahwa ESD = $W_{(15)}$ = 2,63862 diberikan oleh data No. 15.
5. Dengan n = 15 dan tingkat signifikansi 5%, Tabel 8.3 memberikan titik kritis eksak C = 2,65. Nah, karena ESD < C maka berdasarkan uji Grubbs, data No. 15 dinyatakan bukan *outlier*.

Gambar 8.1. Diagram kotak untuk data pada Tabel 8.1

Pada gambar ini, tanda asterisks * berada di bagian atas. Ia menunjukkan data ekstrim kanan 297 layak dijadikan tersangka *outlier*. Sementara itu, di bagian bawah tidak tampak adanya tanda tersebut. Ini berarti, dengan menggunakan teknik Tukey, data ekstrim kiri 165 tidak layak dijadikan tersangka *outlier*.

Tabel 8.2. Nilai statistik terurut W

No. Urut	No. Observasi	W Terurut
1	10	0,01726
2	9	0,04963
3	8	0,17908

4	7	0,34090
5	11	0,37111
6	6	0,43710
7	12	0,50057
8	5	0,50273
9	4	0,59982
10	3	0,69691
11	13	0,75948
12	2	0,89110
13	14	1,08313
14	1	1,63548
15	15	2,63662

** Nilai ESD = 2,63662 diberikan oleh observasi No. 15*

Tabel 8.3. Titik kritis untuk ESD pada tingkat signifikansi 0,5%; 1%; dan 5%

n	**Tingkat signifikansi ***		
	0,5%	**1%**	**5%**
10	2,58	2,55	2,39
11	2,66	2,62	2,45
12	2,77	2,71	2,50
13	2,95	2,84	2,57
14	2,96	2,86	2,62
15	3,02	2,91	2,65
16	3,07	2,95	2,70
17	3,13	3,03	2,75
18	3,17	3,08	2,79
19	3,19	3,10	2,80

20	3,41	3,14	2,83
25	3,52	3,34	2,99
30	3,54	3,35	3,03
35	3,61	3,41	3,09
40	3,68	3,51	3,15
45	3,81	3,57	3,17
50	-	3,68	3,23

** Titik kritis ESD yang diberikan oleh Rosner (1975)*

Catatan:
Dari uraian di atas dapat kita simpulkan bahwa, menurut uji Grubbs (ESD), tidak ada *outlier* di dalam kelompok 15 data itu. Baiklah. Kesimpulan ini dapat kita terima secara logis. Akan tetapi, apakah kesimpulan tersebut cukup bijak (*wise*) mengingat titik kritis yang digunakan dalam analisis di atas diperoleh dari kajian simulasi?

Pada Bab 9 yang akan segera tiba, kita akan menyaksikan bahwa kesimpulan tersebut tidak bijak tatkala kita berpijak pada titik kritis yang eksak.

2. Komputasi ESD Dengan MS Excel
Berikut adalah cara meghitung ESD dengan menggunakanMS Excel.

1. Masuk kedalam sistem MS EXCEL
2. Simpan data di kolom A, dimulai dari baris pertama A1 sampai dengan A15 (bila n = 15)
3. Simpan kursor di kolom A17. Lalu ketik kalimat (tanpa tanda apostrof): “=AVERAGE(A1:A15)”
4. Klik ENTER. Maka di A17 akan muncul nilai rata-rata data
5. Simpan kursor di kolom A18. Lalu ketik (kalimat antara “ dan “): “=STDEV(A1:A15)”
6. Klik ENTER. Maka di A18 akan tampak nilai deviasi standar data
7. Letakkan kursor di B1. Lalu ketik (kalimat antara “ dan “): “=(A1-$A$17)/$A$18” dan klik ENTER

8. Letakkan kursor di pojok kanan bawah sel B1 sampai muncul tanda + berwarna hitam. Lalu klik dua kali serentak
9. Maka di kolom B1 – B15 tetera nilai W1 – W15
10. Salinlah semua nilai di kolom B1 – B15 ke kolom D1 – D15. Lalu nilai-nilai yang ada di D1 – D15 disorot (*highlighted*).
11. Klik "Sort Smallest to Largest". Maka di kolom D1 – D15 sekarang tampak nilai-nilai W(1) – W(15)
12. Nilai ESD diberikan di sel D15.

Selamat mencoba!

BAB 9. UJI KESAHIHAN: TEKNIK IESD

"*All great scientists have, in a certain sense, been great artists; the man with no imagination may collect facts, but he cannot make great discoveries.*"

Karl Pearson

1. Apa Itu Uji IESD?

Telah dikemukakan di Bab 1 bahwa proses pembersihan data terdiri atas 3 tahap, yakni (1) proses penyelidikan untuk mengidentifikasi data yang dicurigai sebagai calon tersangka *outlier*, (2) proses penyidikan untuk mendeteksi apakah data yang dijadikan calon tersangka *outlier* layak atau tidak layak ditetapkan sebagai tersangka, dan (3) proses menguji apakah tersangka sahih (signifikan) atau tidak sahih sebagai *outlier*.

Tahap pertama telah dibahas pada Bab 3 dan Bab 4. Sementara Bab 5 dan Bab 6 diperuntukkan bagi tahap kedua. Lalu, Bab 7 menyajikan teknik pertama untuk melaksanakan tahap ketiga dengan menggunakan statistik penguji Grubbs (ESD). Teknik ini dipilih mengingat, seperti dikemukakan oleh Tietjen dan Moore (1972) dan Rosner (1975), ia memiliki kuasa (*power of the test*) yang optimal tatkala menguji satu *single outlier*. Namun, walaupun demikian, dalam pandangan kami para penulis ada kelemahan yang dimiliki ESD yang dapat mengakibatkan keputusan yang keliru. Kelemahan itu terletak pada penentuan titik kritis yang ditentukan berdasarkan kajian simulasi (lihat Rosner (1975) dan Iglewicz dan Hoaglin (1993)). Penting untuk dicatat bahwa, secara teoritis, hasil kajian simulasi dapat jauh berbeda dengan nilai yang eksak (sebenarnya).

Mengenai kelemahan ESD tersebut di atas beserta jalan keluarnya, para pembaca disarankan untuk berkonsultasi dengan artikel Djauhari (2001a, 2001b). Pada kedua artikel ini,

1. Ditunjukkan secara matematis bahwa distribusi dari ESD adalah distribusi Beta dengan derajat kebebasan 0,5 dan (n – 2)/2.

2. Diberikan solusi untuk menghitung nilai titik kritis yang eksak dalam bentuk persamaan integral. Lalu, perhitungannya dilakukan dengan menggunakan integrasi Monte Carlo berdasarkan 1.000.000 bilangan acak yang dibangkitkan dari distribusi uniform di interval antara 0 dan 1.
3. Disajikan contoh keunggulan teknik yang diperkenalkan oleh Djauhari (2001a, 2001b) berdasarkan data dari dunia kesehatan. Data pada contoh tersebut (lihat kolom kedua pada Tabel 8.1, Bab 8) akan dianalisis kembali di akhir bab ini.

Untuk selanjutnya, teknik ESD yang dilengkapi dengan nilai titik kritis yang eksak, penulis beri nama teknik "Improved ESD" atau disingkat IESD. Sebagaimana halnya ESD, tentu IESD pun dapat digunakan untuk $n \geq 7$.

2. Komputasi IESD Dengan Bantuan MS Excel

Bagi yang berminat mendapatkan penjelasan teoritis dan analitis secara mendetil tentang teknik IESD, silahkan baca Djauhari (2001a, 2001b) yang dapat diunduh dari akun <Maman Djauhari> di *ResearchGate*. Namun, bagi mereka yang ingin langsung mengunakan IESD dalam praktik, misalnya dengan n = 21, berikut ini adalah langkah-langkahnya dengan menggunakan MS Excel.

1. Masuk kedalam sistem MS Excel
2. Simpan data di kolom A, dimulai dari baris pertama A1 sampai dengan A21
3. Simpan kursor di kolom A23. Lalu ketik (kalimat di antara " dan "): "=AVERAGE(A1:A21)"
4. Klik ENTER. Maka di A23 akan tampak nilai rata-rata data
5. Simpan kursor di kolom A24. Lalu ketik (kalimat antara " dan "): "=STDEV(A1:A21)"
6. Klik ENTER. Maka di A24 akan muncul nilai deviasi standar data
7. Letakkan kursor di B1. Lalu ketik (kalimat di antara " dan "): "=(A1-$A$23)/$A$24" dan klik ENTER
8. Letakkan kursor di pojok kanan bawah sel B1 sampai muncul tanda + berwarna hitam. Lalu klik dua kali serentak
9. Maka di kolom B1 – B21 tertera nilai $W_1 - W_{21}$

10. Salinlah semua nilai di kolom B1 – B21 ke kolom D1 – D21. Lalu nilai-nilai yang ada di D1 – D21 disorot (*highlighted*).
11. Klik "Sort Smallest to Largest". Maka di kolom D1 – D21 sekarang tampak nilai-nilai $W_{(1)} – W_{(21)}$
12. Nilai ESD diberikan di sel D21.
13. Untuk menghitung nilai titik kritis C, letakkan kursor di salah satu sel/kotak kosong mana saja. Lalu ketiklah kalimat berikut (tanpa tanda " dan "); "=BETA.INV((1-P)^(1/n);0,5;(n-2)/2;0;1)" dengan P adalah tingkat signifikansi. Selanjutnya, klik ENTER.
14. Maka nilai C muncul di sel tersebut. Untuk ilustrasi, coba masukkan nilai P = 5% dan ukuran sampel n = 10. Maka akan diperoleh C = 0,645461391.

Sangat mudah, bukan? Bahkan cukup dengan MS Excel saja dan tidak perlu menggunakan perangkat lunak yang besar dan mahal. Ini sesuai dengan falsafah: "Jangan membunuh nyamuk pakai bedil!"

Sekali lagi, yang tertarik dengan pendakian dan penurunan matematisnya, silahkan baca Djauhari (2001a, 2001b). Statistik penguji yang terbit dalam kedua artikel tersebut dimaksudkan untuk menandingi keempat statistik penguji berikut: (1) uji Thompson (1985), (2) uji Irwin (1925), (3) uji Student (yang dapat dilihat di referensi # 18), dan (4) uji Dixon (1950). Informasi dapat diperoleh dengan mengunjungi akun <Maman Djauhari> di *ResearchGate*.

Mengingat kelebihan IESD dibandung ESD, kedua artikel tersebut di atas telah membawa penulis menjadi konsultan bagi peneliti/ilmuwan di berbagai lembaga industri internasional di mancanegara. Sebagai ilustrasi, pada contoh berikut akan ditunjukkan kelebihan IESD tersebut.

Contoh

Kita perhatikan kembali data pada kolom kedua, Tabel 8.1, Bab 8, tentang nilai kolesterol 15 orang yang sehat. Dengan menggunakan ESD di bab tersebut telah kita uji bahwa, secara signifikan, kelompok 15 data itu bersih dari kehadiran outlier. Sekarang akan kita uji lagi apakah kedua data ekstrim 165 dan 297 merupakan outlier dengan menggunakan IESD. Berikut adalah hasil pengujian yang telah dilakukan.

1. Hasil perhitungan nilai statistik W_1, W_2, …, W_{15} dapat dilihat pada Tabel 8.1 kolom ketiga, Bab 8.
2. Kelima belas nilai tersebut lalu diurutkan dari yang terkecil sampai dengan terbesar dan diberi lambang $W_{(1)}$, $W_{(2)}$, …, $W_{(15)}$. Hasilnya tertera pada Tabel 8.2 kolom terakhir, Bab 8. Pada tabel tersebut tampak bahwa nilai maksimum diberikan oleh data No. 15, yakni $W_{(15)}$ = 2,63862.
3. Dengan n = 15 dan tingkat signifikansi 5%, Tabel 9.1 di bawah ini memberikan titik kritis eksak C = 2,54589. Nah, karena nilai maksimum $W_{(15)} > C$, maka menurut IESD data No. 15 dinyatakan sebagai outlier. Perhatikanlah bahwa keputusan ini berbeda dengan yang diberikan oleh ESD.

Tabel 9.1. Titik kritis IESD dan ESD* pada tingkat signifikansi 0,5%; 1%; 5%; dan 10%

N	0,5%		1%		5%		10%	
	IESD	ESD	IESD	ESD	IESD	ESD	IESD	ESD
10	2,55712	2,58	2,48967	2,55	2,28797	2,39	2,16781	TA
11	2,64932	2,66	2,57274	2,62	2,35245	2,45	2,22521	TA
12	2,73017	2,77	2,64507	2,71	2,40923	2,50	2,27580	TA
13	2,80095	2,95	2,71009	2,84	2,45974	2,57	2,32112	TA
14	2,86388	2,96	2,76729	2,86	2,50487	2,62	2,36193	TA
15	2,92176	3,02	2,81910	2,91	2,54589	2,65	2,39895	TA
16	2,97218	3,07	2,86605	2,95	2,58319	2,70	2,43304	TA
17	3,01784	3,13	2,90802	3,03	2,61752	2,75	2,46436	TA
18	3,06282	3,17	2,94720	3,08	2,64899	2,79	2,49348	TA
19	3,10240	3,19	2,98347	3,10	2,67834	2,80	2,52043	TA
20	3,13788	3,41	3,01650	3,14	2,70559	2,83	2,54568	TA
25	3,29078	3,52	3,15398	3,34	2,81894	2,99	2,65128	TA
30	3,40473	3,54	3,25806	3,35	2,90597	3,03	2,73315	TA
35	3,49994	3,61	3,34047	3,41	2,97553	3,09	2,79933	TA
40	3,56698	3,68	3,40542	3,51	3,03327	3,15	2,85500	TA
45	3,62884	3,81	3,46355	3,57	3,08220	3,17	2,90235	TA

50	3,68345	TA	3,51091	3,68	3,12465	3,23	2,94406	TA
60	3,77225	TA	3,58945	TA	3,19566	TA	3,01361	TA
70	3,83806	TA	3,65387	TA	3,25355	TA	3,07030	TA
80	3,89644	TA	3,70478	TA	3,30140	TA	3,11804	TA
90	3,95099	TA	3,74929	TA	3,34270	TA	3,15938	TA
100	3,98871	TA	3,78626	TA	3,37907	TA	3,19542	TA

* *Titik kritis ESD yang diberikan oleh Rosner (1975)*

TA: *Tidak ada*

Dengan mengetahui bahwa data No. 15 berupa outlier, pertanyaannya sekarang adalah: "Apakah data No. 1 juga berupa outlier?" Untuk menjawab pertanyaan ini, data outlier No. 15 tersebut kita keluarkan dari kelompoknya. Jadi, sekarang kita berhadapan dengan 14 buah data acak. Dengan demikian, pertanyaannya menjadi: "Apakah data No. 1 berupa outlier di antara n = 14 data sisa itu?" Silahkan pembaca kerjakan sendiri sebagai Latihan. Dengan menggunakan metode yang sama seperti di atas, maka akan diperoleh nilai maksimum $W_{(14)}$ = 2,03849. Selanjutnya, Tabel 9.1 memberikan titik kritis eksaknya C = 2,50487 untuk tingkat signifikansi 5%. Nah, ternyata $W_{(14)} < C$ yang berarti data No. 1 bukan merupakan outlier.

Sekarang kita simpulkan bahwa di dalam kelompok 15 data pada Tabel 8.1 di Bab 8, terdapat satu buah outlier yakni data No. 15 yang bernilai 297.

Sebagai penutup bab ini, perlu dicatat bahwa:

1. Pada Tabel 9.1 di atas, nilai ESD diberikan hanya sampai dua angka di belakang koma desimal sesuai dengan aslinya yang ada di Grubbs (1950). Sedangkan untuk nilai IESD, banyaknya digit di belakang koma desimal dapat disesuaikan dengan tujuan riset.
2. Seperti halnya ESD, IESD dapat digunakan untuk $n \geq 7$. Untuk ukuran sampel n yang sangat kecil (misalnya n = 4 atau 5), statistik penguji Dixon sangat tepat digunakan sebagaimana yang diutarakan oleh Iglewitz dan Hoaglin (1993). Ukuran sampel

seperti ini biasa dijumpai dalam riset di mana data sangat mahal dan/atau lama untuk diperoleh; seperti riset dalam bidang farmaseutikal.

3. Untuk n yang besar, bahkan sampai orde ratusan ribu, gunakanlah statistik penguji tangguh (*robust*) yang kami sebut FMV (*Fast Minimum Variance*) di Bab 13. Ukuran sampel seperti ini biasa digunakan oleh para peneliti bidang sains sosial.
4. FMV adalah hal khusus dari FMCD (*Fast Minimum Covariance Determinant*) yang sangat terkenal dan juga dari MVV (*Minimum Vector Variance*) yang saling komplementer dengan FMCD. Apabila FMCD dan MVV dibangun untuk menguji kehadiran data *outlier* pada kasus multivariat secara sekaligus (bukan satu-per-satu), FMV tidak lain adalah prosedur FMCD dan MVV yang diterapkan untuk kasus univariat.
5. Pengujian hipotesis keadiran *outlier* dengan menggunakan teknik FMCD dan teknik MVV akan dikemukakan dalam buku tersendiri mengenai pembersihan data multivariat.

BAB 10. UJI KESAHIHAN: TEKNIK DIXON

"*Research is what I'm doing when I don't know what I'm doing*."
Wernher von Braun

Dibandingkan dengan teknik Grubbs (ESD) yang dibahas pada Bab 9, teknik IESD lebih ungul. Keunggulannya terletak pada distribusi probabilitas statistik penguji yang dapat dikenalpasti secara eksak. Distribusi eksak inilah yang mampu memberikan nilai titik kritis yang eksak. Namun, di samping keunggulan tersebut, ada keterbatasan yang dimiliki IESD. Apa keterbatasannya? Seperti halnya ESD, tidak ada yang merekomendasikan penggunaannya tatkala ukuran sampel n kecil. Di dalam literatur tentang pengujian hipotesis kehadiran *outlier* untuk data univariat, statistik penguji yang selalu direkomendasikan penggunaannya jika n kecil adalah uji Dixon (Dixon (1950)).

1. Uji Dixon

Ada enam jenis statistik penguji Dixon; 3 jenis untuk menguji bila hanya ada 1 tersangka *outlier*, dan 3 jenis lagi untuk 2 tersangka. Keenam jenis tersebut dapat dilihat pada Tabel 10.1.

Tabel 10.1. Enam jenis statistik penguji Dixon

No.	Simbol	Statistik Penguji	Data yang diuji	Ukuran sampel
1	r_{10}	$[X_{(n)}-X_{(n-1)}]/[X_{(n)}-X_{(1)}]$	$X_{(n)}$	$3 \le n \le 7$
2	r_{11}	$[X_{(n)}-X_{(n-1)}]/[X_{(n)}-X_{(2)}]$ atau $[X_{(2)}-X_{(1)}]/[X_{(n-1)}-X_{(1)}]$	$X_{(n)}$ $X_{(1)}$	$8 \le n \le 10$
3	r_{12}	$[X_{(n)}-X_{(n-1)}]/[X_{(n)}-X_{(3)}]$ atau $[X_{(2)}-X_{(1)}]/[X_{(n-2)}-X_{(1)}]$	$X_{(n)}$ $X_{(1)}$	$8 \le n \le 10$
4	r_{20}	$[X_{(n)}-X_{(n-2)}]/[X_{(n)}-X_{(1)}]$ atau $[X_{(3)}-X_{(1)}]/[X_{(n)}-X_{(1)}]$	$X_{(n)}$ dan $X_{(n-1)}$ $X_{(1)}$ dan $X_{(2)}$	$11 \le n \le 13$

5	r_{21}	$[X_{(n)}-X_{(n-2)}]/[X_{(n)}-X_{(2)}]$ atau $[X_{(3)}-X_{(1)}]/[X_{(n-1)}-X_{(1)}]$	$X_{(n)}$ dan $X_{(n-1)}$ $X_{(1)}$ dan $X_{(2)}$	$11 \leq n \leq 13$
6	r_{22}	$[X_{(n)}-X_{(n-2)}]/[X_{(n)}-X_{(3)}]$ atau $[X_{(3)}-X_{(1)}]/[X_{(n-2)}-X_{(1)}]$	$X_{(n)}$ dan $X_{(n-1)}$ $X_{(1)}$ dan $X_{(2)}$	$14 \leq n \leq 30$

Sekali lagi, $X_{(1)}$ adalah data dengan nilai terkecil, $X_{(2)}$ terkecil kedua, $X_{(3)}$ terkecil ketiga, …, $X_{(n)}$ data terbesar.

Secara umum, teknik Dixon digunakan untuk n yang kecil antara 3 dan 30. Uraian yang lebih terperinci dikemukakan oleh Manoj dan Kannan (2013) seperti yang tertera di kolom terakhir Tabel 10.1. Di situ terpampang sebagai berikut.

1. Uji r_{10} untuk $3 \leq n \leq 7$
2. Uji r_{11} untuk $8 \leq n \leq 10$; Uji r_{12} identik dengan r_{11}
3. Uji r_{21} untuk $11 \leq n \leq 13$; Uji r_{20} identik dengan r_{21}
4. Uji r_{22} untuk $14 \leq n \leq 30$.

Dengan menggunakan statistik penguji Dixon yang mana saja, hipotesis H0: "Tidak ada outlier dalam kelompok data" ditolak bila nilai statistik penguji tersebut lebih besar dari pada titik kritisnya.

Perhatikan dengan seksama Tabel 10.1 di atas. Tabel itu hanya menyediakan fasilitas untuk menguji sekaligus pasangan 2 data terbesar (2 ekstrim kanan) atau 2 data terkecil (2 ekstrim kiri). Dengan kata lain, tabel itu tidak menyediakan statistik penguji untuk menguji sekaligus untuk pasangan ekstrim kiri $X_{(1)}$ dan ekstrim kanan $X_{(n)}$. Kendala lain yang dihadapi oleh para pengguna statistik tatkala menggunakan uji Dixon adalah pada penentuan titik kritis yang didasarkan kepada hasil kajian simulasi. Walaupun sangatlah wajar apabila titik kritis untuk teknik Dixon ditentukan melalui simulasi. Mengapa wajar? Sebab, kalau kita teliti, keenam jenis statistik penguji Dixon dibangun atas dasar perbedaan yang mencolok antara data ekstrim dengan data terdekatnya. Nah, di dalam statistika, cara pandang seperti ini akan memudahkan kita membuat formula statistik

penguji, tapi menyulitkan kita menurunkan distribusi probabilitasnya. Dengan demikian, simulasi adalah solusi tatkala distribusi probabilitas tidak diketahui.

Selain kendala itu, perlu dicatat pula bahwa teknik Dixon ini tidak dicantumkan sebagai bahan kajian dan diskusi dalam *Handbook of Engineering Statistics* (referensi # 18); sebuah buku pegangan para pengguna statistika di bidang rekayasa. Tampaknya teknik Dixon kurang mendapat tempat di kalangan pengguna statistika rekayasa (*engineering statistics*).

2. Titik Kritis

Pada dasarnya teknik Dixon dapat digunakan untuk data yang berasal dari populasi yang berdistribusi apa saja asalkan distribusi itu dapat diidentifikasi. Inilah kelebihan teknik Dixon. Namun, kendalanya adalah pada penentuan titik kritis. Sebagai ilustrasi, andaikan kita berhadapan dengan n buah data acak yang berasal dari populasi berdistribusi tertentu. Berikut adalah cara menentukan titik kritis melalui kajian simulasi.

1. Bangkitkan data acak sebanyak n dari distribusi tersebut di atas.
2. Berdasarkan n buah data itu, hitunglah r_{10}, r_{11}, r_{12}, r_{20}, r_{21}, dan r_{22}.
3. Ulangi Langkah 1 dan Langkah 2 di atas sebanyak N kali. Seperti dikemukakan Tietjen-Moore (1972), banyaknya iterasi N yang khas (*typical*) adalah N = 10.000. Namun, Ruiz and Verma (2006, halaman 136) menggunakan N = 100.000.
4. Dengan N = 100.000, maka untuk masing-masing r_{10}, r_{11}, r_{12}, r_{20}, r_{21}, dan r_{22} diperoleh 100.000 buah data acak hasil simulasi.
5. Hitunglah persentil ke-$(1 - \alpha)$% dari 100.000 buah data acak tersebut baik untuk r_{10}, r_{11}, r_{12}, r_{20}, r_{21}, maupun r_{22} dengan α = 0,005; 0,01; 0,02; 0,05; 0,10; 0,20; dan 0,30.
6. Hasil hitungan inilah yang dijadikan titik kritis untuk r_{10}, r_{11}, r_{12}, r_{20}, r_{21}, dan r_{22} dengan ukuran sampel n dan tingkat signifikansi 0,5%; 1%; 2%; 5%; 10%; 20%; dan 30%.

Pada Tabel 10.2 dan Tabel 10.3 di bawah ini disajikan titik kritis untuk r_{10}, r_{11}, r_{12}, dan untuk r_{20}, r_{21}, r_{22} pada berbagai ukuran sampel n di mana data acak sebanyak n berasal dari populasi berdistribusi normal dengan tingkat signifikansi 1% dan 5%. Tabel yang lebih rinci diberikan di Apendiks. Sebagai contoh, titik kritis untuk r_{10} dengan n = 5 dan tingkat signifikansi 5% adalah C = 0.6423 yang dapat dilihat pada Tabel 10.2, kolom ketiga, baris ketiga.

Tabel 10.2. Titik kritis untuk r_{10}, r_{11} dan r_{12} dengan tingkat signifikansi 1% dan 5%

n	r_{10}		r_{11}		r_{12}	
	1%	5%	1%	5%	1%	5%
3	0,9881	0,9411				
4	0,8886	0,7651	0,9910	0,9550		
5	0,7819	0,6423	0,9120	0,8064	0,9920	0,9597
6	0,6987	0,5624	0,8185	0,6916	0,9220	0,8248
7	0,6371	0,5077	0,7399	0,6111	0,8347	0,7152
8	0,5914	0,4673	0,6808	0,5539	0,7618	0,6365
9	0,5554	0,4363	0,6346	0,5114	0,7047	0,5798
10	0,5260	0,4122	0,5972	0,4778	0,6587	0,5361
11	0,5028	0,3922	0,5663	0,4510	0,6207	0,5018
12	0,4831	0,3755	0,5412	0,4291	0,5903	0,4740
13	0,4664	0,3615	0,5208	0,4111	0,5651	0,4519
14	0,4517	0,3496	0,5026	0,3955	0,5431	0,4327
15	0,4385	0,3389	0,4868	0,3819	0,5243	0,4162
16	0,4268	0,3293	0,4723	0,3698	0,5076	0,4015
17	0,4166	0,3208	0,4595	0,3594	0,4929	0,3889
18	0,4081	0,3135	0,4495	0,3500	0,4807	0,3779
19	0,4002	0,3068	0,4395	0,3418	0,4692	0,3682
20	0,3922	0,3005	0,4303	0,3340	0,4588	0,3590
21	0,3854	0,2947	0,4220	0,3271	0,4493	0,3511
22	0,3789	0,2895	0,4143	0,3207	0,4404	0,3436

23	0,3740	0,2851	0,4081	0,3151	0,4329	0,3369
24	0,3674	0,2804	0,4006	0,3092	0,4244	0,3302
25	0,3625	0,2763	0,3949	0,3043	0,4179	0,3245
26	0,3583	0,2725	0,3893	0,2995	0,4117	0,3191
27	0,3543	0,2686	0,3848	0,2954	0,4062	0,3143
28	0,3499	0,2655	0,3795	0,2912	0,4007	0,3095
29	0,3460	0,2622	0,3748	0,2874	0,3952	0,3052
30	0,3425	0,2594	0,3702	0,2837	0,3905	0,3012

Sumber: Ruiz dan Verma (2006)

Tabel 10.3. Titik kritis untuk r_{20}, r_{21} dan r_{22} dengan tingkat signifikansi 1% dan 5%

n	r_{20}		r_{21}		r_{22}	
	1%	5%	1%	5%	1%	5%
4	0,9934	0,9669				
5	0,9294	0,8445	0,9952	0,9760		
6	0,8458	0,7400	0,9462	0,8778	0,9959	0,9794
7	0,7743	0,6640	0,8756	0,7842	0,9526	0,8919
8	0,7156	0,6077	0,8106	0,7110	0,8890	0,8051
9	0,6694	0,5644	0,7561	0,6545	0,8289	0,7351
10	0,6325	0,5305	0,7111	0,6102	0,7771	0,6802
11	0,6024	0,5028	0,6736	0,5745	0,7336	0,6359
12	0,5769	0,4798	0,6425	0,5454	0,6974	0,5998
13	0,5562	0,4613	0,6174	0,5213	0,6671	0,5707
14	0,5379	0,4448	0,5945	0,5007	0,6403	0,5457
15	0,5213	0,4300	0,5754	0,4823	0,6177	0,5240
16	0,5066	0,4178	0,5573	0,4667	0,5973	0,5054
17	0,4944	0,4066	0,5425	0,4530	0,5798	0,4891
18	0,4831	0,3967	0,5292	0,4409	0,5643	0,4746
19	0,4724	0,3878	0,5168	0,4300	0,5503	0,4621
20	0,4636	0,3797	0,5059	0,4198	0,5381	0,4501
21	0,4547	0,3721	0,4955	0,4109	0,5263	0,4398
22	0,4473	0,3655	0,4866	0,4026	0,5162	0,4305
23	0,4401	0,3591	0,4784	0,3952	0,5065	0,4218
24	0,4331	0,3534	0,4701	0,3879	0,4969	0,4133
25	0,4273	0,3478	0,4633	0,3814	0,4893	0,4058
26	0,4215	0,3429	0,4564	0,3751	0,4819	0,3989
27	0,4160	0,3382	0,4498	0,3700	0,4744	0,3927
28	0,4108	0,3336	0,4438	0,3644	0,4676	0,3866

29	0,4057	0,3296	0,4381	0,3595	0,4617	0,3812
30	0,4014	0,3256	0,4330	0,3550	0,4556	0,3759

Sumber: Ruiz dan Verma (2006)

3. Komputasi

Komputasi statistik penguji Dixon didominasi oleh proses mengurutkan data dari data terkecil sampai dengan data terbesar. Oleh karena itu, cara mengurutkan data dengan bantuan MINITAB yang kita bahas pada Bab 4 diperlukan lagi di sini. Untuk membangunkan ingatan kita, berikut ini kita tulis kembali langkah-langkahnya.

1. Masuk kedalam sistem MINITAB
2. Simpan nomor observasi di kolom C1, dimulai dari baris ke-1
3. Simpan data di kolom C2, dimulai dari baris ke-1
4. Simpan kursor di kolom C3 baris pertama
5. Klik ikon "Data"
6. Klik "Sort…"
7. Pada window "Column to sort by" tulis "C2" dibawah "Column"
8. Klik OK
9. Pada kolom C3 tampak nomor obervasi sesuai dengan data yang sudah diurutkan
10. Pada kolom C4 kita peroleh data terurut dari data terkecil $X_{(1)}$, data terkecil kedua $X_{(2)}$, …, sampai dengan yang terbesar $X_{(n)}$.
11. Hitunglah nilai statistik penguji yang digunakan sesuai rumus pada Tabel 10.1.

Contoh

Perhatikan kembali data pada Tabel 8.1, Bab 8. Dengan menggunakan teknik Dixon, apakah kedua data ekstrim 165 dan 297 merupakan *outlier*?

Untuk menjawab pertanyaan ini, mari kita gunakan $r_{11} = [X_{(n)}-X_{(n-1)}]/[X_{(n)}-X_{(2)}]$ yang memungkinkan kita dapat menguji data ekstrim

kanan maupun data ekstrim kiri. Untuk menguji bahwa 297 merupakan *outlier*, kita hitung dulu nilai r11. Dari data pada Tabel 8.1 pada Bab 8 itu, kita peroleh r_{11} = (297 – 249)/(297 – 188) = 0,44037. Adapun titik kritisnya untuk tingkat signifikansi 5%, dengan n = 15, adalah C = 0,3819 yang dapat dibaca pada Tabel 10.3 kolom kelima dan baris keduabelas. Karena $r_{11} > C$, maka berdasarkan uji Dixon, data 297 dinyatakan sebagai *outlier*.

Selanjutnya, untuk menguji apakah 165 berupa *outlier* atau bukan, kita keluarkan 297 dari kelompok data itu. Jadi, sekarang kita berhadapan dengan n = 14. Kita uji 165 dengan menggunakan $r_{11} = [X_{(2)}-X_{(1)}]/[X_{(n-1)}-X_{(1)}]$. Kali ini, data pada Tabel 8.1, Bab 8, itu memberikan r_{11} = (188 – 165)/(249 – 165) = 0,27381. Sedangkan titik kritisnya, untuk tingkat signifikansi 5% dan n = 14, adalah C = 0,3955 (lihat Tabel 10.3, kolom kelima, baris kesebelas, di atas). Ternyata, kali ini $r_{11} < C$ yang berarti 165 bukan *outlier*.

Dari analisis di atas dapat kita simpulkan bahwa hasil yang diberikan oleh teknik Dixon sama dengan hasil dari teknik IESD. Kedua teknik pengujian ini memperkuat argumen yang dikemukakan oleh teknik Tukey bahwa data ekstrim kanan 297 layak dijadikan tersangka *outlier*.

Sebagai penutup bab ini, perlu dicatat bahwa tatkala n kecil dan hanya ada satu tersangka *outlier*, penggunaan teknik Dixon sangat populer di kalangan para praktisi.

APENDIKS

Keenam tabel pada Apendiks ini, dari Tabel 10.4 sampai dengan Tabel 10.9, menyajikan titik kritis untuk r_{10}, r_{11}, r_{12}, r_{20}, r_{21}, dan r_{22} dengan tingkat signifikansi 0,5%, 1%, 2%, 5%, 10%, 20% dan 30% pada berbagai ukuran sampel n yang berasal dari distribusi normal. Semua tabel tersebut bersumber dari Ruiz dan Verma (2006).

Tabel 10.4. Titik kritis untuk r_{10}

n	Tingkat signifikansi						
	0,5%	**1%**	**2%**	**5%**	**10%**	**20%**	**30%**
3	0,9940	0,9881	0,9763	0,9411	0,8850	0,7808	0,6836
4	0,9201	0,8886	0,8457	0,7651	0,6789	0,5603	0,4704
5	0,8234	0,7819	0,7291	0,6423	0,5578	0,4508	0,3730
6	0,7437	0,6987	0,6458	0,5624	0,4840	0,3868	0,3173
7	0,6809	0,6371	0,5864	0,5077	0,4340	0,3444	0,2811
8	0,6336	0,5914	0,5432	0,4673	0,3979	0,3138	0,2550
9	0,5952	0,5554	0,5091	0,4363	0,3704	0,2915	0,2361
10	0,5658	0,5260	0,4813	0,4122	0,3492	0,2735	0,2208
11	0,5416	0,5028	0,4591	0,3922	0,3312	0,2586	0,2086
12	0,5208	0,4831	0,4405	0,3755	0,3170	0,2467	0,1983
13	0,5034	0,4664	0,4250	0,3615	0,3045	0,2366	0,1898
14	0,4869	0,4517	0,4118	0,3496	0,2938	0,2280	0,1826
15	0,4739	0,4385	0,3991	0,3389	0,2848	0,2202	0,1764
16	0,4614	0,4268	0,3883	0,3293	0,2765	0,2137	0,1707
17	0,4504	0,4166	0,3792	0,3208	0,2691	0,2077	0,1656
18	0,4423	0,4081	0,3711	0,3135	0,2626	0,2023	0,1613
19	0,4333	0,4002	0,3630	0,3068	0,2564	0,1973	0,1572
20	0,4247	0,3922	0,3562	0,3005	0,2511	0,1929	0,1535
21	0,4173	0,3854	0,3495	0,2947	0,2460	0,1890	0,1504
22	0,4109	0,3789	0,3439	0,2895	0,2415	0,1854	0,1474
23	0,4051	0,3740	0,3384	0,2851	0,2377	0,1820	0,1446

24	0,3986	0,3674	0,3328	0,2804	0,2337	0,1790	0,1420
25	0,3935	0,3625	0,3287	0,2763	0,2303	0,1761	0,1397
26	0,3889	0,3583	0,3242	0,2725	0,2269	0,1735	0,1376
27	0,3843	0,3543	0,3202	0,2686	0,2237	0,1710	0,1355
28	0,3801	0,3499	0,3163	0,2655	0,2208	0,1687	0,1335
29	0,3762	0,3460	0,3127	0,2622	0,2182	0,1664	0,1318
30	0,3718	0,3425	0,3093	0,2594	0,2155	0,1645	0,1300

Tabel 10.5. Titik kritis untuk r_{11}

n	Tingkat signifikansi						
	0,5%	**1%**	**2%**	**5%**	**10%**	**20%**	**30%**
4	0,9955	0,9910	0,9820	0,9550	0,9105	0,8226	0,7351
5	0,9376	0,9120	0,8762	0,8064	0,7281	0,6147	0,5244
6	0,8544	0,8185	0,7722	0,6916	0,6098	0,5020	0,4208
7	0,7812	0,7399	0,6918	0,6111	0,5332	0,4328	0,3589
8	0,7226	0,6808	0,6321	0,5539	0,4793	0,3858	0,3177
9	0,6757	0,6346	0,5876	0,5114	0,4404	0,3519	0,2883
10	0,6375	0,5972	0,5509	0,4778	0,4102	0,3260	0,2660
11	0,6055	0,5663	0,5215	0,4510	0,3857	0,3054	0,2483
12	0,5796	0,5412	0,4977	0,4291	0,3660	0,2886	0,2341
13	0,5581	0,5208	0,4780	0,4111	0,3496	0,2747	0,2223
14	0,5396	0,5026	0,4605	0,3955	0,3357	0,2632	0,2125
15	0,5229	0,4868	0,4456	0,3819	0,3236	0,2530	0,2039
16	0,5080	0,4723	0,4322	0,3698	0,3129	0,2443	0,1965
17	0,4945	0,4595	0,4204	0,3594	0,3037	0,2365	0,1900
18	0,4845	0,4495	0,4102	0,3500	0,2952	0,2297	0,1843
19	0,4734	0,4395	0,4010	0,3418	0,2878	0,2235	0,1791
20	0,4639	0,4303	0,3926	0,3340	0,2810	0,2178	0,1744
21	0,4551	0,4220	0,3847	0,3271	0,2747	0,2128	0,1703
22	0,4472	0,4143	0,3776	0,3207	0,2692	0,2083	0,1664
23	0,4406	0,4081	0,3714	0,3151	0,2644	0,2041	0,1630
24	0,4325	0,4006	0,3646	0,3092	0,2593	0,2002	0,1596
25	0,4267	0,3949	0,3591	0,3043	0,2550	0,1965	0,1567
26	0,4208	0,3893	0,3537	0,2995	0,2509	0,1933	0,1540
27	0,4158	0,3848	0,3493	0,2954	0,2472	0,1901	0,1514
28	0,4107	0,3795	0,3444	0,2912	0,2435	0,1872	0,1489
29	0,4053	0,3748	0,3403	0,2874	0,2404	0,1845	0,1466
30	0,4007	0,3702	0,3362	0,2837	0,2371	0,1821	0,1446

Tabel 10.6. Titik kritis untuk r_{12}

n	Tingkat signifikansi						
	0,5%	**1%**	**2%**	**5%**	**10%**	**20%**	**30%**
5	0,9960	0,9920	0,9839	0,9597	0,9195	0,8381	0,7551
6	0,9448	0,9220	0,8895	0,8248	0,7503	0,6399	0,5503
7	0,8683	0,8347	0,7918	0,7152	0,6356	0,5281	0,4450
8	0,8005	0,7618	0,7151	0,6365	0,5588	0,4572	0,3814
9	0,7447	0,7047	0,6575	0,5798	0,5048	0,4091	0,3385
10	0,6984	0,6587	0,6116	0,5361	0,4645	0,3735	0,3073
11	0,6603	0,6207	0,5752	0,5018	0,4329	0,3461	0,2836
12	0,6290	0,5903	0,5455	0,4740	0,4076	0,3242	0,2646
13	0,6031	0,5651	0,5215	0,4519	0,3868	0,3065	0,2496
14	0,5803	0,5431	0,5006	0,4327	0,3697	0,2921	0,2370
15	0,5613	0,5243	0,4827	0,4162	0,3547	0,2793	0,2263
16	0,5435	0,5076	0,4665	0,4015	0,3418	0,2685	0,2170
17	0,5285	0,4929	0,4523	0,3889	0,3304	0,2590	0,2090
18	0,5161	0,4807	0,4406	0,3779	0,3203	0,2507	0,2019
19	0,5041	0,4692	0,4298	0,3682	0,3116	0,2433	0,1958
20	0,4931	0,4588	0,4199	0,3590	0,3035	0,2365	0,1901
21	0,4833	0,4493	0,4106	0,3511	0,2962	0,2306	0,1851
22	0,4742	0,4404	0,4025	0,3436	0,2897	0,2252	0,1804
23	0,4664	0,4329	0,3952	0,3369	0,2839	0,2202	0,1765
24	0,4573	0,4244	0,3874	0,3302	0,2781	0,2156	0,1725
25	0,4507	0,4179	0,3814	0,3245	0,2731	0,2114	0,1690
26	0,4437	0,4117	0,3752	0,3191	0,2683	0,2076	0,1659
27	0,4383	0,4062	0,3702	0,3143	0,2641	0,2039	0,1628
28	0,4325	04007	0,3645	0,3095	0,2597	0,2005	0,1599
29	0,4266	0,3952	0,3600	0,3052	0,2560	0,1973	0,1573
30	0,4219	0,3905	0,3552	0,3012	0,2524	0,1946	0,1550

Tabel 10.7. Titik kritis untuk r_{20}

n	Tingkat signifikansi						
	0,5%	**1%**	**2%**	**5%**	**10%**	**20%**	**30%**
4	0,9967	0,9934	0,9867	0,9669	0,9345	0,8703	0,8069
5	0,9496	0,9294	0,9005	0,8445	0,7822	0,6934	0,6230
6	0,8770	0,8458	0,8070	0,7400	0,6734	0,5863	0,5205
7	0,8091	0,7743	0,7324	0,6640	0,5990	0,5164	0,4556
8	0,7514	0,7156	0,6741	0,6077	0,5451	0,4671	0,4102
9	0,7055	0,6694	0,6289	0,5644	0,5048	0,4305	0,3770
10	0,6674	0,6325	0,5929	0,5305	0,4731	0,4021	0,3511
11	0,6367	0,6024	0,5635	0,5028	0,4472	0,3793	0,3304
12	0,6106	0,5769	0,5391	0,4798	0,4264	0,3607	0,3135
13	0,5888	0,5562	0,5188	0,4613	0,4090	0,3449	0,2994
14	0,5702	0,5379	0,5012	0,4448	0,3937	0,3316	0,2873
15	0,5536	0,5213	0,4857	0,4300	0,3803	0,3199	0,2768
16	0,5383	0,5066	0,4714	0,4178	0,3690	0,3098	0,2676
17	0,5256	0,4944	0,4598	0,4066	0,3587	0,3008	0,2596
18	0,5133	0,4831	0,4492	0,3967	0,3498	0,2927	0,2524
19	0,5029	0,4724	0,4392	0,3878	0,3414	0,2855	0,2458
20	0,4931	0,4636	0,4301	0,3797	0,3340	0,2788	0,2400
21	0,4840	0,4547	0,4221	0,3721	0,3272	0,2728	0,2347
22	0,4763	0,4473	0,4151	0,3655	0,3210	0,2673	0,2297
23	0,4687	0,4401	0,4082	0,3591	0,3152	0,2623	0,2253
24	0,4614	0,4331	0,4018	0,3534	0,3100	0,2577	0,2211
25	0,4555	0,4273	0,3956	0,3478	0,3048	0,2533	0,2172
26	0,4495	0,4215	0,3902	0,3429	0,3005	0,2494	0,2137
27	0,4433	0,4160	0,3853	0,3382	0,2961	0,2455	0,2104
28	0,4381	0,4108	0,3803	0,3336	0,2919	0,2420	0,2074
29	0,4331	0,4057	0,3757	0,3296	0,2882	0,2388	0,2043
30	0,4285	0,4014	0,3714	0,3256	0,2845	0,2356	0,2015

Tabel 10.8. Titik kritis untuk r_{21}

n	Tingkat signifikansi						
	0,5%	**1%**	**2%**	**5%**	**10%**	**20%**	**30%**
5	0,9976	0,9952	0,9905	0,9760	0,9519	0,9019	0,8501
6	0,9620	0,9462	0,9236	0,8778	0,8248	0,7461	0,6802
7	0,9012	0,8756	0,8427	0,7842	0,7236	0,6408	0,5763
8	0,8411	0,8106	0,7735	0,7110	0,6493	0,5682	0,5068
9	0,7886	0,7561	0,7176	0,6545	0,5942	0,5162	0,4579
10	0,7444	0,7111	0,6722	0,6102	0,5515	0,4763	0,4209
11	0,7068	0,6736	0,6354	0,5745	0,5172	0,4451	0,3915
12	0,6754	0,6425	0,6050	0,5454	0,4893	0,4197	0,3684
13	0,6497	0,6174	0,5801	0,5213	0,4668	0,3988	0,3491
14	0,6268	0,5945	0,5583	0,5007	0,4472	0,3810	0,3330
15	0,6073	0,5754	0,5392	0,4823	0,4302	0,3659	0,3192
16	0,5887	0,5573	0,5221	0,4667	0,4158	0,3528	0,3072
17	0,5734	0,5425	0,5076	0,4530	0,4028	0,3413	0,2967
18	0,5594	0,5292	0,4948	0,4409	0,3917	0,3310	0,2874
19	0,5468	0,5168	0,4829	0,4300	0,3814	0,3218	0,2791
20	0,5360	0,5059	0,4721	0,4198	0,3721	0,3135	0,2716
21	0,5247	0,4955	0,4624	0,4109	0,3639	0,3061	0,2649
22	0,5160	0,4866	0,4538	0,4026	0,3561	0,2992	0,2586
23	0,5075	0,4784	0,4456	0,3952	0,3492	0,2930	0,2530
24	0,4982	0,4701	0,4377	0,3879	0,3426	0,2872	0,2479
25	0,4918	0,4633	0,4311	0,3814	0,3364	0,2818	0,2431
26	0,4845	0,4564	0,4244	0,3751	0,3310	0,2769	0,2386
27	0,4778	0,4498	0,4184	0,3700	0,3259	0,2723	0,2346
28	0,4718	0,4438	0,4127	0,3644	0,3209	0,2680	0,2306
29	0,4657	0,4381	0,4076	0,3595	0,3164	0,2641	0,2271
30	0,4605	0,4330	0,4024	0,3550	0,3122	0,2602	0,2237

Tabel 10.9. Titik kritis untuk r_{22}

n	Tingkat signifikansi						
	0,5%	**1%**	**2%**	**5%**	**10%**	**20%**	**30%**
6	0,9980	0,9959	0,9918	0,9794	0,9582	0,9140	0,8668
7	0,9665	0,9526	0,9327	0,8919	0,8438	0,7697	0,7068
8	0,9122	0,8890	0,8591	0,8051	0,7477	0,6679	0,6042
9	0,8573	0,8289	0,7942	0,7351	0,6756	0,5962	0,5346
10	0,8084	0,7771	0,7405	0,6802	0,6207	0,5431	0,4842
11	0,7654	0,7336	0,6965	0,6359	0,5775	0,5023	0,4457
12	0,7296	0,6974	0,6599	0,5998	0,5425	0,4697	0,4153
13	0,6990	0,6671	0,6299	0,5707	0,5148	0,4436	0,3908
14	0,6722	0,6403	0,6039	0,5457	0,4907	0,4214	0,3704
15	0,6494	0,6177	0,5817	0,5240	0,4702	0,4029	0,3534
16	0,6286	0,5973	0,5617	0,5054	0,4528	0,3868	0,3386
17	0,6108	0,5798	0,5447	0,4891	0,4372	0,3728	0,3258
18	0,5951	0,5643	0,5299	0,4746	0,4238	0,3606	0,3145
19	0,5804	0,5503	0,5162	0,4621	0,4118	0,3495	0,3046
20	0,5682	0,5381	0,5036	0,4501	0,4010	0,3397	0,2956
21	0,5562	0,5263	0,4925	0,4398	0,3913	0,3309	0,2876
22	0,5455	0,5162	0,4825	0,4305	0,3821	0,3228	0,2802
23	0,5359	0,5065	0,4735	0,4218	0,3742	0,3156	0,2736
24	0,5260	0,4969	0,4643	0,4133	0,3666	0,3089	0,2675
25	0,5186	0,4893	0,4566	0,4058	0,3596	0,3026	0,2619
26	0,5102	0,4819	0,4492	0,3989	0,3533	0,2970	0,2568
27	0,5028	0,4744	0,4425	0,3927	0,3474	0,2916	0,2520
28	0,4956	0,4676	0,4362	0,3866	0,3415	0,2867	0,2475
29	0,4891	0,4617	0,4301	0,3812	0,3365	0,2821	0,2434
30	0,4835	0,4556	0,4245	0,3759	0,3317	0,2777	0,2394

BAB 11. UJI KESAHIHAN: TEKNIK TIETJEN-MOORE

"*Science is the belief in the ignorance of experts*."

Richard P. Feynman

Proses mengenalpasti *outlier* amat sangat kompleks. Justru kompleksitas inilah yang membuat riset tentang pengujian *outlier* dan cara menanganinya menjadi sangat menantang. Mengapa demikian? Karena topik ini membuka peluang yang sangat luas bagi para peneliti untuk turut serta berkontribusi dalam menciptakan teknik-teknik baru dalam penyelidikan, penyidikan, dan pengujian kesahihan data *outlier*.

Pada tiga bab terakhir kita telah membahas tiga teknik; teknik ESD (Grubbs, 1950), teknik IESD (Djauhari, 2001a, 2001b, 2003), dan teknik Dixon (1950). Dua Teknik pertama, ESD dan IESD memiliki kuasa uji yang tinggi untuk menguji kehadiran satu (*single*) *outlier*. Sementara itu, teknik Dixon sangat populer untuk menguji kehadiran satu (*single*) *outlier*. Teknik Dixon pun memberikan fasilitas untuk menguji sekaligus dua data ekstrim kanan atau dua ekstrim kiri. Pengujian sekaligus dimaksudkan untuk menghindari adanya *masking effect* atau *swamping effect*.

Masking effect terjadi kalau kedua tersangka saling melindungi sehingga mereka tidak terdeteksi sebagai *outlier* padahal mereka adalah *outlier*. Ibarat koruptor berjamaah yang saling melindungi satu sama lain. Sementara itu, *swamping effect* adalah keadaan sebaliknya; kedua tersangka adalah *inliers* (data bersih), namun terdeteksi sebagai *outlier*. Sama persis dengan kasus Sengkon-Karta; dua tersangka pembunuhan yang fenomenal dan bahkan boleh dikatakan monumental dalam sejarah peradilan di Indonesia. Mereka sempat dipenjara selama sekitar 4 tahun sebelum pengadilan membebaskan padahal mereka tidak bersalah dan bukan pembunuh (https://www.kompasiana.com/234/54ff676ca33311494c50ff70/legenda-sengkon-dan-karta). Kasus-kasus seperti itu tidak akan pernah terjadi kalau penyusun hipotesis (polisi dan jaksa) dan statistik pengujinya (hakim) berkarakter tangguh (*robust*).

Apabila teknik ESD dan teknik IESD melibatkan proses penaksiran mean dan standar deviasi populasi, tidak demikian halnya dengan teknik Dixon yang hanya meibatkan data ekstrim. Akibatnya, teknik Dixon bersifat tangguh. Inilah antara lain yang membuat banyak orang gemar menggunakan teknik tersebut. Namun, teknik Dixon hanya cocok digunakan untuk ukuran sampel n yang kecil. Sementara itu, walaupun tidak tangguh, IESD dapat digunakan untuk n yang moderat atau besar, $n \geq 7$. Namun demikian, IESD hanya untuk menguji satu tersangka.

Pertanyaannya sekarang adalah: "Apabila terdapat k buah tersangka dengan $k > 2$, bagaimana mengujinya secara sekaligus?" Pengujian sekaligus dilakukan untuk menghindari adanya *masking effect* dan/atau *swamping effect*. Untuk menjawab pertanyaan ini, kami perkenalkan teknik Tietjen-Moore yang dapat digunakan untuk menguji sekaligus kehadiran k buah *outliers* dengan k boleh = 1 atau 2 atau … berapa saja. Hipotesisnya sebagai berikut.

H0: Tidak ada *outlier* di dalam sekelompok data
H1: Ada tepat *k* buah *outlier* dalam sekelompok data

1. Cara Kerja Teknik Tietjen-Moore

Di bawah ini dikemukakan langkah-langkah praktis pengujian dengan menggunakan teknik Tietjen-Moore. Bagi mereka yang ingin mengetahui teorinya secara detil, artikel Tietjen-Moore (1972) adalah referensi aslinya.

Berikut ini langkah-langkah praktisnya. Pertama, urutkan data dari yang terkecil nilainya hingga yang terbesar. Sebagaimana telah dikemukakan di Bab 10, data yang sudah diurutkan itu kita beri lambang $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$ di mana $X_{(1)}$ data terkecil, $X_{(2)}$ terkecil kedua, …, $X_{(n)}$ terbesar. Kemudian,

1. Untuk menguji k buah data ekstrim kanan, hitunglah statistik penguji $L_{ka} = (n-1)*SS_k/s^2$ di mana
   1) s = deviasi standar data
   2) $SS_k = (X_{(1)}-X_{bark})^2 + (X_{(2)}-X_{bark})^2 + \ldots + (X_{(n-k)}-X_{bark})^2$, dengan X_{bark} = rata-rata data tanpa k buah data terbesar

2. Untuk menguji k buah data ekstrim kiri, hitung statistik penguji $L_{ki} = (n-1)*SS_k/s^2$ di mana
   1) s = deviasi standar data
   2) $SS_k = (X_{(k+1)}-X_{bark})^2 + (X_{(k+2)}-X_{bark})^2 + \ldots + (X_{(n)}-X_{bark})^2$. Di sini X_{bark} = rata-rata data tanpa k buah data terkecil
3. Bila k buah data yang mau diuji berada di kedua ekstrim (kiri dan kanan), hitung statistik penguji $E_k = (n-1)*SS_k/S^2$ di mana,
   1) S = deviasi standar data $Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)}$
   2) $SS_k = (Z_{(1)}-Z_{bark})^2 + (Z_{(2)}-Z_{bark})^2 + \ldots + (Z_{(n-k)}-Z_{bark})^2\}$, di mana Z_{bark} = rata-rata dari $Z_{(1)}, Z_{(2)}, \ldots, Z_{(n-k)}$
   3) $Z_{(1)}$ adalah data terkecil di antara $Z_1, Z_2, \ldots, Z_n$; $Z_{(2)}$ adalah data yang terkecil kedua, $Z_{(3)}$ adalah data yang terkecil ketiga, …, $Z_{(n)}$ adalah data yang terbesar.
   4) $Z_1 = ABS(X_{(1)}-X_{bar})$, $Z_2 = ABS(X_{(2)}-X_{bar})$, …, $Z_n = ABS(X_{(n)}-X_{bar})$, dan
   5) X_{bar} = rata-rata dari $X_1, X_2, \ldots, X_n$.

Jadi, ada tiga jenis statistik penguji pada teknik Tietjen-Moore yakni (1) L_{ka} untuk menguji k buah data ekstrim kanan, (2) L_{ki} untuk menguji k buah data ekstrim kiri, dan (3) E_k untuk menguji k buah data yang berada di kedua ekstrim (kiri dan kanan). Nilai ketiga statistik penguji ini terletak antara 0 dan 1. Nilai tersebut akan dekat dengan 1 jika H0 benar; artinya tidak ada *outlier* di dalam data yang kita analisis. Dan, akan dekat dengan 0 apabila ada *outlier*.

Dengan begitu, pengambilan keputusannya adalah sebagai berikut. Hipotesis H0 ditolak apabila nilai statistik penguji Tietjen-Moore kecil (daerah penolakan ada di sebelah kiri). Dengan kata lain, H0 ditolak pada tingkat signifikansi tertentu dengan titik kritis C, jika nilai statistik penguji Tietjen-Moore < C.

Dalam kasus di mana populasi berdistribusi normal, untuk k = 1, teknik Tietjen-Moore tidak lain adalah teknik IESD (lihat referensi # 18). Namun, karena titik kritis pada teknik IESD adalah eksak (ditentukan melalui distribusi probabilitas), sedangkan titik kritis pada teknik Tietjen-Moore ditentukan melalui simulasi, dalam kasus k = 1 ini penggunaan teknik IESD lebih kami anjurkan.

2. Titik Kritis

Titik kritis pada teknik Tietjen-Moore ditentukan melalui simulasi dengan cara membangkitkan sampel acak dari disribusi normal standar berukuran n. Untuk keperluan ini, biasanya digunakan n = 10.000. Nilai statistik penguji Tietjen-Moore yang diperoleh dari data kemudian dibandingkan dengan titik kritis tersebut.

Seperti telah dikemukakan di bagian pertama di atas, nilai statistik penguji Tietjen-Moore berada di antara 0 dan 1. Jika tidak ada *outlier* dalam data, nilai tersebut mendekati 1. Sebaliknya, jika ada *outlier*, nilai itu akan mendekati 0. Dengan demikian, teknik Tietjen-Moore selalu merupakan pengujian satu sisi dengan daerah penolakan di sebelah kiri terlepas dari statistik penguji mana yang digunakan, L_{ka}, L_{ki}, atau E_k.

Pada Tabel 11.1 sampai dengan Tabel 11.6 disajikan nilai titik kritis dari ketiga statistik penguji itu untuk berbagai nilai k dan berbagai tingkat signifikansi.

Tabel 11.1(a). Titik kritis L_{ka} dan L_{ki} dengan tingkat signifikansi 1% dan k < 6

n	k						
	1	**1***	**2**	**2****	**3**	**4**	**5**
4	0,011	0,010					
5	0,045	0,044	0,004	0,004			
6	0,091	0,093	0,021	0,019	0,002		
7	0,148	0,145	0,047	0,044	0,010		
8	0,202	0,195	0,076	0,075	0,028	0,008	
9	0,235	0,241	0,112	0,108	0,048	0,018	
10	0,280	0,283	0,142	0,142	0,070	0,032	0,012
11	0,327	0,321	0,178	0,174	0,098	0,052	0,026
12	0,371	0,355	0,208	0,204	0,120	0,070	0,038
13	0,400	0,386	0,233	0,233	0,147	0,094	0,056
14	0,424	0,414	0,267	0,261	0,172	0,113	0,072

15	0,450	0,440	0,294	0,286	0,194	0,132	0,090
16	0,473	0,463	0,311	0,310	0,219	0,151	0,108
17	0,480	0,485	0,338	0,332	0,237	0,171	0,126
18	0,502	0,504	0,358	0,353	0,260	0,192	0,140
19	0,508	0,522	0,366	0,373	0,272	0,201	0,154
20	0,533	0,539	0,387	0,391	0,300	0,231	0,175
25	0,603	0,607	0,468		0,377	0,308	0,246
30	0,650		0,526		0,434	0,369	0,312
35	0,690		0,574		0,484	0,418	0,364
40	0,722		0,608		0,522	0,460	0,408
45	0,745		0,636		0,558	0,498	0,444
50	0,768		0,668		0,592	0,531	0,483

** Dari Grubbs* (*1950, Table I*)
*** Dari Grubbs* (*1950, Table V*)

Tabel 11.1(b). Titik kritis L_{ka} dan L_{ki} dengan tingkat signifikansi 1% dan k > 5

n	k				
	6	**7**	**8**	**9**	**10**
12	0,019				
13	0,033				
14	0,046	0,026			
15	0,057	0,037			
16	0,072	0,049	0,030		
17	0,091	0,064	0,044		
18	0,104	0,076	0,053	0.036	
19	0,118	0,088	0,064	0,046	
20	0,136	0,104	0,078	0,058	0,042
25	0,204	0,168	0,144	0,112	0,092

30	0,268	0,229	0,196	0,166	0,142
35	0,321	0,282	0,250	0,220	0,194
40	0,364	0,324	0,292	0,262	0,234
45	0,399	0,361	0,328	0,296	0,270
50	0,438	0,400	0,368	0,336	0,308

Tabel 11.2(a). Titik kritis L_{ka} dan L_{ki} dengan tingkat signifikansi 5% dan k < 6

n	k						
	1	**1***	**2**	**2****	**3**	**4**	**5**
3	0,003	0,003					
4	0,051	0,049	0,001	0,001			
5	0,125	0,127	0,018	0,018			
6	0,203	0,203	0,055	0,057	0,010		
7	0,273	0,270	0,106	0,102	0,032		
8	0,326	0,326	0,146	0,148	0,064	0,022	
9	0,372	0,374	0,194	0,191	0,099	0,045	
10	0,418	0,415	0,233	0,230	0,129	0,070	0,034
11	0,454	0,451	0,270	0,267	0,162	0,098	0,054
12	0,489	0,482	0,305	0,300	0,196	0,125	0,076
13	0,517	0,510	0,337	0,330	0,224	0,150	0,098
14	0,540	0,534	0,363	0,357	0,250	0,174	0,122
15	0,556	0,556	0,387	0,382	0,276	0,197	0,140
16	0,575	0,576	0,410	0,405	0,300	0,219	0,159
17	0,594	0,593	0,427	0,426	0,322	0,240	0,181
18	0,608	0,610	0,447	0,446	0,337	0,259	0,200
19	0,624	0,624	0,462	0,464	0,354	0,277	0,209
20	0,639	0,638	0,484	0,480	0,377	0,299	0,238

25	0,696	0,692	0,550	0,450	0,374	0,312
30	0,730		0,599	0,506	0,434	0,376
35	0,762		0,642	0,554	0,482	0,424
40	0,784		0,672	0,588	0,523	0,468
45	0,802		0,696	0,618	0,556	0,502
50	0,820		0,722	0,646	0,588	0,535

* *Dari Grubbs* (*1950, Table I*)
** *Dari Grubbs* (*1950, Table V*)

Tabel 11.2(b). Titik kritis L_{ka} dan L_{ki} dengan tingkat signifikansi 5% dan k > 5

n	k				
	6	**7**	**8**	**9**	**10**
12	0,042				
13	0,060				
14	0,079	0,050			
15	0,097	0,066			
16	0,115	0,082	0,055		
17	0,136	0,100	0,072		
18	0,154	0,116	0,086	0,062	
19	0,168	0,130	0,099	0,074	
20	0,188	0,150	0,115	0,088	0,066
25	0,262	0,222	0,184	0,154	0,126
30	0,327	0,283	0,245	0,212	0,183
35	0,376	0,334	0,297	0,264	0,235
40	0,421	0,378	0,342	0,310	0,280
45	0,456	0,417	0,382	0,350	0,320
50	0,490	0,450	0,414	0,383	0,356

Tabel 11.3(a). Titik kritis L_{ka} dan L_{ki} dengan tingkat signifikansi 10% dan k < 6

n	k						
	1	**1***	**2**	**2****	**3**	**4**	**5**
3	0,011	0,011					
4	0,098	0,098	0,003	0,003			
5	0,200	0,199	0,038	0,038			
6	0,280	0,283	0,091	0,092	0,020		
7	0,348	0,350	0,148	0,148	0,056		
8	0,404	0,405	0,200	0,199	0,095	0,038	
9	0,448	0,450	0,248	0,245	0,134	0,068	
10	0,490	0,488	0,287	0,286	0,170	0,098	0,051
11	0,526	0,520	0,326	0,323	0,208	0,128	0,074
12	0,555	0,548	0,361	0,355	0,240	0,159	0,103
13	0,578	0,573	0,388	0,384	0,270	0,186	0,126
14	0,600	0,594	0,416	0,411	0,298	0,212	0,150
15	0,611	0,613	0,436	0,435	0,322	0,236	0,172
16	0,631	0,631	0,458	0,456	0,342	0,260	0,194
17	0,648	0,646	0,478	0,476	0,364	0,282	0,216
18	0,661	0,660	0,496	0,494	0,384	0,302	0,236
19	0,676	0,673	0,510	0,511	0,398	0,316	0,251
20	0,688	0,685	0,530	0,527	0,420	0,339	0,273
25	0,732	0,732	0,588		0,489	0,412	0,350
30	0,766		0,637		0,523	0,472	0,411
35	0,792		0,673		0,586	0,516	0,458
40	0,812		0,702		0,622	0,554	0,499
45	0,826		0,724		0,648	0,586	0,533
50	0,840		0,744		0,673	0,614	0,562

* *Dari Grubbs* (*1950, Table I*)
** *Dari Grubbs* (*1950, Table V*)

Tabel 11.3(b). Titik kritis L_{ka} dan L_{ki} dengan tingkat signifikansi 10% dan k > 5

n	k				
	6	**7**	**8**	**9**	**10**
12	0,062				
13	0,082				
14	0,104	0,068			
15	0,124	0,086			
16	0,144	0,104	0,073		
17	0,165	0,125	0,092		
18	0,184	0,142	0,108	0,080	
19	0,199	0,158	0,124	0,094	
20	0,220	0,176	0,140	0,110	0,085
25	0,296	0,251	0,213	0,189	0,152
30	0,359	0,316	0,276	0,240	0,210
35	0,410	0,365	0,328	0,294	0,262
40	0,451	0,408	0,372	0,338	0,307
45	0,488	0,447	0,410	0,378	0,348
50	0,518	0,477	0,442	0,410	0,380

Tabel 11.4. Titik kritis E_k dengan tingkat signifikansi 1%

n	k									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
4	0,004									
5	0,029	0,002								
6	0,068	0,012	0,001							
7	0,110	0,028	0,006							
8	0,156	0,050	0,014	0,004						
9	0,197	0,078	0,026	0,009						
10	0,235	0,101	0,018	0,006						
11	0,274	0,134	0,064	0,030	0,012					
12	0,311	0,159	0,083	0,042	0,020	0,008				
13	0,337	0,181	0,103	0,056	0,031	0,014				
14	0,374	0,207	0,123	0,072	0,042	0,022	0,012			
15	0,404	0,238	0,146	0,090	0,054	0,032	0,018			
16	0,422	0,263	0,166	0,107	0,068	0,040	0,024	0,014		
17	0,440	0,290	0,188	0,122	0,079	0,052	0,032	0,018		
18	0,459	0,306	0,206	0,141	0,094	0,062	0,041	0,026	0,014	
19	0,484	0,323	0,219	0,156	0,108	0,074	0,050	0,032	0,020	
20	0,499	0,339	0,236	0,170	0,121	0,086	0,058	0,040	0,026	0,017
25	0,571	0,418	0,320	0,245	0,188	0,146	0,110	0,087	0,066	0,050
30	0,624	0,482	0,386	0,308	0,250	0,204	0,166	0,132	0,108	0,087
35	0,669	0,533	0,435	0,364	0,299	0,252	0,211	0,177	0,149	0,124
40	0,704	0,574	0,480	0,408	0,347	0,298	0,258	0,220	0,190	0,164
45	0,728	0,607	0,518	0,446	0,386	0,336	0,294	0,258	0,228	0,200
50	0,748	0,636	0,550	0,482	0,424	0,376	0,334	0,297	0,264	0,235

Tabel 11.5. Titik kritis E_k dengan tingkat signifikansi 5%

n	k										
	1	**1***	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
3	0,001	0,001									
4	0,025	0,025	0,001								
5	0,081	0,081	0,010								
6	0,146	0,145	0,034	0,004							
7	0,208	0,207	0,065	0,016							
8	0,265	0,262	0,099	0,034	0,010						
9	0,314	0,310	0,137	0,057	0,021						
10	0,356	0,352	0,172	0,083	0,037	0,014					
11	0,386	0,390	0,204	0,107	0,055	0,026					
12	0,424	0,423	0,234	0,133	0,073	0,039	0,018				
13	0,455	0,453	0,262	0,156	0,092	0,053	0,028				
14	0,484	0,479	0,293	0,179	0,112	0,068	0,039	0,021			
15	0,509	0,503	0,317	0,206	0,134	0,084	0,052	0,030			
16	0,526	0,525	0,340	0,227	0,153	0,102	0,067	0,041	0,024		
17	0,544	0,544	0,362	0,248	0,170	0,116	0,078	0,050	0,032		
18	0,562	0,562	0,382	0,267	0,187	0,132	0,091	0,062	0,041	0,026	
19	0,581	0,579	0,390	0,287	0,203	0,146	0,105	0,074	0,050	0,033	
20	0,597	0,594	0,416	0,302	0,221	0,163	0,119	0,085	0,059	0,041	0,028
25	0,652	0,654	0,493	0,381	0,298	0,236	0,186	0,146	0,114	0,089	0,068
30	0,698		0,549	0,443	0,364	0,298	0,246	0,203	0,166	0,137	0,112
35	0,732		0,596	0,495	0,417	0,351	0,298	0,254	0,214	0,181	0,154
40	0,758		0,629	0,534	0,458	0,395	0,343	0,297	0,259	0,223	0,195
45	0,778		0,658	0,567	0,492	0,433	0,381	0,337	0,299	0,263	0,233
50	0,797		0,684	0,599	0,529	0,468	0,417	0,373	0,334	0,299	0,268

** Dari Grubbs* (*1950, Table I*).

Tabel 11.6. Titik kritis E_k dengan tingkat signifikansi 10%

n	k										
	1	**1***	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
3	0,003	0,003									
4	0,050	0,049	0,002								
5	0,127	0,127	0,022								
6	0,204	0,203	0,056	0,009							
7	0,268	0,270	0,094	0,027							
8	0,328	0,326	0,137	0,053	0,016						
9	0,377	0,374	0,175	0,080	0,032						
10	0,420	0,415	0,214	0,108	0,052	0,022					
11	0,449	0,451	0,250	0,138	0,073	0,036					
12	0,485	0,482	0,278	0,162	0,094	0,052	0,026				
13	0,510	0,510	0,309	0,189	0,116	0,068	0,038				
14	0,538	0,534	0,337	0,216	0,138	0,086	0,052	0,029			
15	0,558	0,556	0,360	0,240	0,160	0,105	0,067	0,040			
16	0,578	0,576	0,384	0,263	0,182	0,122	0,082	0,053	0,032		
17	0,594	0,593	0,406	0,284	0,198	0,140	0,095	0,064	0,042		
18	0,610	0,610	0,424	0,304	0,217	0,156	0,110	0,076	0,051	0,034	
19	0,629	0,624	0,442	0,322	0,234	0,172	0,124	0,089	0,062	0,042	
20	0,644	0,638	0,460	0,338	0,252	0,188	0,138	0,102	0,072	0,051	0,035
25	0,693	0,692	0,528	0,417	0,331	0,264	0,210	0,168	0,132	0,103	0,080
30	0,730		0,582	0,475	0,391	0,325	0,270	0,224	0,186	0,154	0,126
35	0,763		0,624	0,523	0,443	0,379	0,324	0,276	0,236	0,202	0,172
40	0,784		0,657	0,562	0,486	0,422	0,367	0,320	0,284	0,243	0,212
45	0,803		0,684	0,593	0,522	0,459	0,406	0,360	0,320	0,284	0,252
50	0,820		0,708	0,622	0,552	0,492	0,440	0,396	0,355	0,319	0,287

* *Dari Grubbs* (*1950, Table I*).

Contoh

Data pada Tabel 11.7, kolom kedua, disajikan kembali 15 buah data yang digunakan oleh Grubbs (1950). Data itu digunakan lagi oleh Tietjen-Moore (1972) sebagai contoh untuk memberikan ilustrasi bagaimana statistik penguji E_k bekerja. Mari kita analisis lebih lanjut.

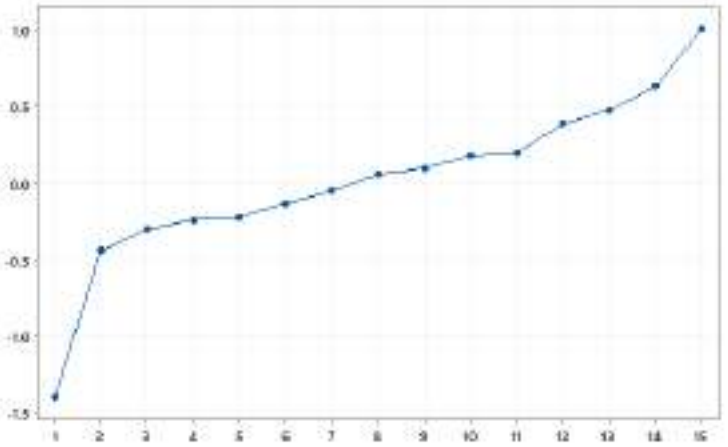
Tabel 11.7. Data dan berbagai statistik untuk menghitung E_k dengan k = 2

No.	Data	r_i	r_i Terurut	Z	Z-Z_{bar}	Z-Z_{bark}
1	-1,40	1,418	0,042	0,06	0,042	0,009
2	-0,44	0,458	0,068	-0,05	-0,068	-0,101
3	-0,30	0,318	0,082	0,10	0,082	0,049
4	-0,24	0,258	0,148	-0,13	-0,148	-0,181
5	-0,22	0,238	0,162	0,18	0,162	0,129
6	-0,13	0,148	0,182	0,20	0,182	0,149
7	-0,05	0,068	0,238	-0,22	-0,238	-0,271
8	0,06	0,042	0,258	-0,24	-0,258	-0,291
9	0,10	0,082	0,318	-0,30	-0,318	-0,351
10	0,18	0,162	0,372	0,39	0,372	0,339
11	0,20	0,182	0,458	-0,44	-0,458	-0,491
12	0,39	0,372	0,462	0,48	0,462	0,429
13	0,48	0,462	0,612	0,63	0,612	0,579
14	0,63	0,612	0,992	1,01	0,992	
15	1,01	0,992	1,418	-1,40	-1,418	
MEAN	0,018					
Z_{bar}				0,018		
Z_{bark}				0,051		
SUMSQ*					4,250	1,241

* SUMSQ: *sum of squares*

Pertama kita tunjukkan cara kerja E_k seperti yang ditunjukkan Tietjen-Moore (1972). Begini caranya,

1. Run chart pada Gambar 11.1 menunjukkan bahwa data terkecil (–1,40) patut dicurigai sebagai calon tersangka *outlier*.
2. Tietjen-Moore (1972) berani mengatakan: "This plot indicates that the normality assumption is reasonable." Bahkan dia menulis: "To a lesser extent, the maximum value may also be an outlier." Oleh karena itu, Tietjen-Moore langsung menguji hipotesis bahwa kedua data ekstrim kiri dan kanan (–1,40 dan 1,01) adalah *outlier*.



Gambar 11.1. Diagram deretan 15 data Grubbs (1950)

3. Karena kedua data ekstrim itu berada pada ekstrim kiri dan ekstrim kanan, maka Tietjen-Moore (1972) menggunakan uji E_k.
4. Oleh karena itulah, yang pertama harus kita hitung adalah rata-rata data (*sample mean*) X_{bar}. Pada Tabel 11.7 itu, hasilnya adalah MEAN = 0,018. Lalu, kita hitung statistik r_i sebagai berikut (hasilnya disajikan pada Tabel 11.7, kolom ketiga):

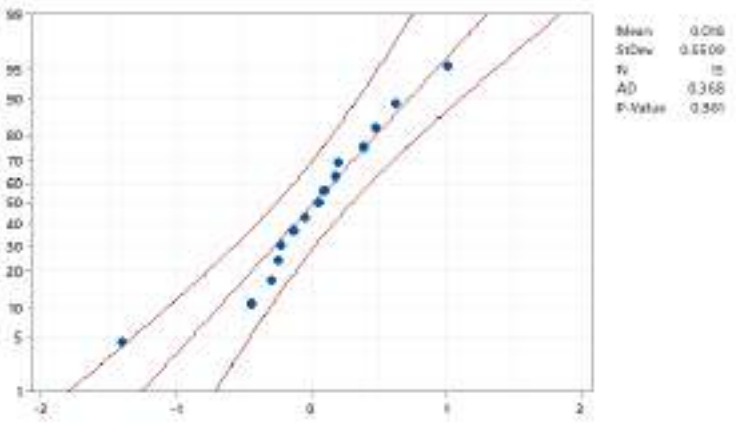
$r_1 = ABS(X_1-X_{bar})$, $r_2 = ABS(X_2-X_{bar})$, …, $r_{15} = ABS(X_{15}-X_{bar})$

5. Selanjutnya nilai r_1, r_2, …, r_{15} diurutkan dari yang terkecil sampai dengan terbesar. Hasilnya tertera pada kolom keempat pada Tabel 11.7.
6. Kemudian, kelimabelas data diurutkan sesuai dengan urutan nilai r_i pada Langkah 5. Kelimabelas data terurut ini diberi lambang Z (kolom kelima pada Tabel 11.7).
7. Sekarang, kita hitung Z_{bar}; rata-rata semua data Z. Tentu saja, Z_{bar} = MEAN = 0,018. Kita hitung pula Z_{bark} rata-rata semua data Z tanpa kedua data yang dicurigai sebagai *outlier*, yakni rata-rata dari Z_1, Z_2, …, Z_{13}. Hasilnya adalah Z_{bark} = 0,051.
8. Tahap berikutnya adalah memusatkan data Z. Artinya, setiap data Z dari nomor 1 sampai dengan nomor 15, dikurangi Z_{bar}. Hasilnya ada pada kolom keenam.
9. Lalu, setiap data Z dari nomor 1 sampai dengan nomor 13 (tanpa melibatkan kedua data ekstrim yang dicurigai sebagai *outlier*), dikurangi Z_{bark}. Hasilnya ada pada kolom ketujuh.
10. Tahap terakhir adalah menghitung jumlah kuadrat data Z terpusat pada kolom keenam (hasilnya SUMSQ = 4,250) dan menghitung jumlah kuadrat data Z terpusat pada kolom terakhir (hasilnya SUMSQ = 1,241).
11. Jadi, E_k = 1,241/4,250 = 0,292.
12. Untuk n = 15 dan k = 2, dengan tingkat signifikansi 5%, Tabel 11.5, kolom keempat dan baris ketigabelas, memberikan titik kritis C = 0,317. Karena E_k < C, maka H0 ditolak yang berarti kedua data ekstrim itu berupa outlier.

Begitulah cara kerja statistik penguji E_k dari Langkah 1 – Langkah 11. Lalu pada Langkah 12 dilanjutkan dengan analisis dan diakhiri dengan pengujian hipotesis berdasarkan teknik Tietjen-Moore.

3. Analisis Lebih Lanjut

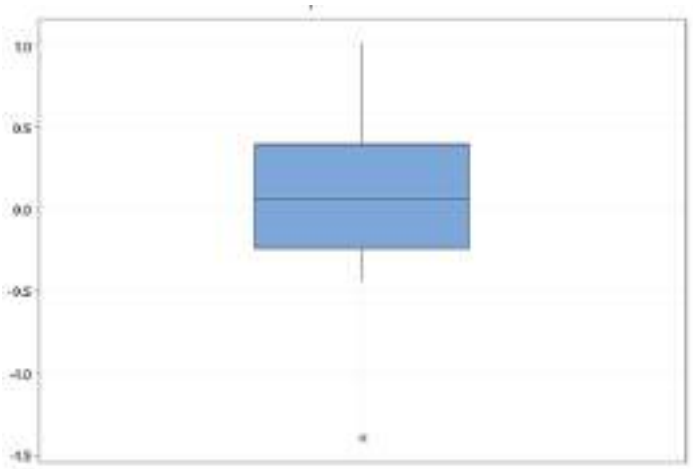
Sekarang, marilah kita lakukan analisis lebih lanjut. Kita mulai dengan melakukan identifikasi melalui *run chart* pada Gambar 11.1 di atas. Berdasarkan gambar itu kita setuju dengan apa yang dikatakan Tietjen-Moore bahwa data terkecil (–1,40) patut dicurigai sebagai tersangka *outlier*. Namun tentang status data terbesar (1,01) akan kita selidiki dengan bantuan teknik Tukey (Bab 5) dan teknik Iglewicz-Hoaglin (Bab 6). Begitu juga dengan kenormalan data.



Gambar 11.2. Diagram probabilitas normal dengan daerah konfidensi 95%

Nilai p-Value 0,381 pada gambar ini menunjukkan kenormalan data dengan pengecualian pada data terkecil yang berada sedikit di luar daerah konfidensi. Sementara itu, data terbesar tampak seperti data lainnya berada di dalam daerah konfidensi. Dengan kata lain, kita kekurangan bukti untuk menjadikan data terbesar ini sebagai tersangka *outlier*. Oleh karena itu, kita memerlukan dukungan alat yang lain untuk memperkuat alasan penetapan sebagai tersangka. Untuk itu akan kita gunakan teknik diagram kotak (Bab 5).

Ternyata, hasil identifikasi dengan menggunakan teknik diagram kotak memperkuat apa yang dihasilkan oleh Gambar 11.2 di atas seperti tampak pada Gambar 11.3.



Gambar 11.3. Teknik diagram kotak untuk mengidentifikasi calon tersangka *outlier*

Pada gambar ini, data terkecil diberi tanda asterisks yang berarti data ini patut dijadikan tersangka. Nah, sekarang kita makin yakin bahwa data terkecil adalah tersangka outlier. Sedangkan data terbesar tidak dapat dijadikan tersangka. Namun demikian, hasil identifikasi ini masih memerlukan analisis yang lebih tajam dengan menggunakan teknik deteksi Tukey dan teknik deteksi Iglewicz-Hoaglin.

3.1. Analisis Berdasarkan Teknik Tukey

Untuk ini kita hitung dulu sari numerik MIN, Q1, MED, Q3 dan MAX. Lalu, kita hitung IQR, BA dan BB dengan konstanta pengali Tukey K = 1,5. Hasilnya tertera pada Tabel 11.8.

Tabel 11.8. Nilai sari numerik

Sari numerik	Nilai
MIN	-1,40
Q1	-0,23
MED	0,06
Q3	0,295

MAX	1,01
IQR	0,525
BA	1,0825
BB	-1,0175

Pada tabel ini tampak BA lebih besar dari data terbesar (1,01) yang berarti data ini tidak layak dijadikan tersangka *outlier*. Adapun data terkecil (–1,40) lebih kecil dari BB. Artinya, sama seperti analisis Tietjen-Moore, ia adalah tersangka *outlier*.

3.2. Analisis Berdasarkan Teknik Iglewicz-Hoaglin
Kita hitung dulu IH-Score untuk setiap data,

$$\text{IH-Score} = 0{,}67449*(X_k - \text{MED})/\text{MAD}$$

dengan k = 1, 2, …, 15 dan MED adalah median semua data. Sedangkan MAD (*median absolute deviation*) diberikan oleh,

$$\text{MAD} = \text{Median dari ABS}(X_1\text{-MED}), \text{ABS}(X_2\text{-MED}), \ldots, \text{ABS}(X_n\text{-MED}).$$

Hasil perhitungan disajikan pada Tabel 11.9 di mana ABSDEV pada kolom ketiga menunjukkan nilai $\text{ABS}(X_1\text{-MED})$, $\text{ABS}(X_2\text{-MED})$, …, $\text{ABS}(X_n\text{-MED})$ untuk setiap data X_1, X_2, …, X_{15}.

Tabel 11.9. Statistik MED, MAD dan IH-Score

No.	Data	ABSDEV	IH-Score
1	-1.40	1.46	-3.28257
2	-0.44	0.50	-1.12417
3	-0.30	0.36	-0.80940
4	-0.24	0.30	-0.67450
5	-0.22	0.28	-0.62953
6	-0.13	0.19	-0.42718

7	-0.05	0.11	-0.24732
8	0.06	0.00	0
9	0.10	0.04	0.08993
10	0.18	0.12	0.26980
11	0.20	0.14	0.31477
12	0.39	0.33	0.74195
13	0.48	0.42	0.94430
14	0.63	0.57	1.28155
15	1.01	0.95	2.13592
MED	0.06		
MAD		0.30	

Pada tabel ini tampak tidak ada data yang lebih besar dari ambang batas 3,5 yang disarankan oleh Iglewicz-Hoaglin. Dengan kata lain, menurut teknik IH tidak ada *outlier*; data itu bersih dari kehadiran *outlier*.

Catatan:
Jangan lupa bahwa penggunaan ambang batas 3,5 diatas ekivalen dengan mengatakan bahwa "banyaknya *outlier* yang ditolerir dalam sekumpulan data adalah tidak lebih dari 0,05%." Kalau ingin dibuat kebijakan yang sama seperti dalam teknik Tukey yakni: "dalam sekumpulan data ada tidak lebih dari 0,7% data yang dapat ditolerir," maka ambang batasnya bukan 3,5 tapi 2,7. Dengan demikian, BA = 2,7 dan BB = – 2,7. Nah, untuk ambang batas 2,7 kita peroleh hasil yang sama seperti yang diberikan oleh teknik Tukey. Data terkecil (– 1,40) memiliki IH-Score = – 3.28257 yang lebih kecil dari BB yang berarti data tersebut layak dijadikan tersangka *outlier*. Di lain pihak, data terbesar (1,01) tidak layak dijadikan tersangka *outlier* sebab memiliki IH-Score = 2.13592 yang lebih kecil dari BA.

3.3. Pengujian Hipotesis
Berdasarkan hasil analisis di atas, sekarang kita lakukan pengujian hipotesis berikut,

H0: Tidak ada *outlier* dalam data
H1: Data terkecil adalah *outlier*

dengan mengggunakan statistik penguji E_k. Hasil perhitungannya disajikan pada Tabel 11.10 yang sama dengan Tabel 11.7 kecuali kolom terakhir.

Tabel 11.10. Data dan berbagai statistik untuk menghitung E_k dengan k = 1

No.	Data	r_i	r_i Terurut	Z	Z-Z_{bar}	Z-Z_{bark}
1	-1,40	1,418	0,042	0,06	0,042	-0,059
2	-0,44	0,458	0,068	-0,05	-0,068	-0,169
3	-0,30	0,318	0,082	0,10	0,082	-0,019
4	-0,24	0,258	0,148	-0,13	-0,148	-0,249
5	-0,22	0,238	0,162	0,18	0,162	0,061
6	-0,13	0,148	0,182	0,20	0,182	0,081
7	-0,05	0,068	0,238	-0,22	-0,238	-0,339
8	0,06	0,042	0,258	-0,24	-0,258	-0,359
9	0,10	0,082	0,318	-0,30	-0,318	-0,419
10	0,18	0,162	0,372	0,39	0,372	0,271
11	0,20	0,182	0,458	-0,44	-0,458	-0,559
12	0,39	0,372	0,462	0,48	0,462	0,361
13	0,48	0,462	0,612	0,63	0,612	0,511
14	0,63	0,612	0,992	1,01	0,992	0,891
15	1,01	0,992	1,418	-1,40	-1,418	
MEAN	0,018					
Z_{bar}				0,018		
Z_{bark}				0,119		
SUMSQ*					4,250	2,095

* SUMSQ: *sum of squares*

Berdasarkan tabel ini kita peroleh E_k = 2,095/4,250 = 0,493. Sedangkan untuk n = 15 dan k = 1, dengan tingkat signifikansi 5%, Tabel 11.5, kolom kedua dan baris ketigabelas, memberikan titik kritis C = 0,509. Nah, karena E_k < C, maka H0 ditolak yang berarti data terbesar sahih sebagai *outlier*.

Sebagai latihan, silakan pembaca menggunakan statistik penguji L_{ki} untuk menguji H0 itu dengan tingkat signifikansi yang sama yakni 5%. Keputusannya akan mengejutkan.

4. Catatan Penutup

Ada perbedaan antara hasil yang diberikan oleh teknik Tietjen-Moore dan hasil analisis yang kita buat di atas. Perbedaan ini menunjukkan kompleksitas masalah pengujian *outlier* yang memerlukan kehati-hatian. Inti perbedaan terletan pada pemilihan hipotesis yang patut dianut di antara kedua hipotesis berikut ini,

H0: Tidak ada *outlier* dalam data
H1: Ada satu *outlier* dalam data

ataukah

H0: Tidak ada *outlier* dalam data
H1: Ada dua *outlier* dalam data

Nah, di sinilah pentingnya mengetahui dengan tepat nilai k (banyaknya *outlier* di dalam data) tatkala menggunakan teknik Tietjen-Moore. Untuk menghadapi masalah penentuan nilai k yang tepat, pada bab berikutnya kita akan membahas teknik Rosner. Namun demikian, perlu dicatat kekhasan dari teknik Tietjen-Moore berikut.

1. Teknik Tietjen-Moore dapat digunakan untuk data yang berasal dari populasi berdistribusi apa saja
2. Teknik ini dapat digunakan untuk menguji kehadiran k buah data *outlier*; k boleh = 1 atau 2 atau … berapa saja
3. Teknik Tietjen-Moore merupakan generalisasi dari teknik Grubbs (ESD). Oleh karena itu, tatkala k = 1 dan data berasal dari populasi normal, lebih baik gunakan saja IESD
4. Teknik ini dapat digunakan untuk $n \geq 7$. Di dalam Tietjen-Moore (1972) diberikan contoh dengan n = 8.

Itulah kelebihan dari teknik Tietjen-Moore. Adapun keterbatasannya:

1. Nilai k harus ditentukan terlebih dahulu dengan tepat
2. Titik kritisnya ditentukan berdasarkan simulasi sesuai dengan distribusi populasi.

BAB 12. UJI KESAHIHAN: TEKNIK ROSNER

"*If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is*."

John von Neumann

Pada empat bab terakhir kita telah membahas empat buah teknik yang sangat populer untuk menguji kehadiran *outlier*. Keempat teknik itu adalah,

1. Teknik Grubbs (ESD) yang diperkenalkan oleh Grubbs (1950)
2. Teknik IESD yang tidak lain adalah uji ESD dengan titik kritis yang eksak (Djauhari, 2001a, 2001b, 2003)
3. Teknik Dixon (r_{10}, r_{11}, r_{12}, r_{20}, r_{21} dan r_{22}) yang dikemukakan oleh Dixon (1950)
4. Teknik Tietjen-Moore (L_{ka}, L_{ki} & E_k) karya Tietjen-Moore (1972).

Tidak lengkap kita membicarakan keempat teknik tersebut tanpa membahas teknik Generalized ESD (GESD) karya Rosner (1983). Seperti telah kita bahas di Bab 11, penggunaan teknik Tietjen-Moore menuntut pengetahuan tentang banyaknya tersangka *outlier* k. Nilai k ini harus diketahui dengan tepat. Ini bukan masalah yang mudah. Untuk mengatasi kesulitan inilah, maka diperkenalkan teknik Rosner. Namun, teknik ini hanya berlaku untuk data yang berasal dari distribusi normal.

1. Apa Itu Teknik Rosner?

Teknik GESD diperkenalkan oleh Rosner (1983) untuk mendeteksi kehadiran 1 atau lebih *outlier* dalam sekelompok data yang berasal dari populasi normal. Jadi, seperti halnya keempat teknik tersebut di atas, teknik Rosner pun bekerja untuk data univariat.

Apa keunggulan teknik GESD? Dibandingkan dengan teknik ESD dan teknik Tietjen-Moore (L_{ka}, L_{ki}, dan E_k), GESD memiliki keistimewaan berikut. Pada teknik Grubbs, banyaknya *outlier* k = 1, sedangkan pada teknik Rosner nilai k boleh lebih dari 1. Selanjutnya,

pada teknik Tietjen-Moore, banyaknya *outlier* di dalam data (k) harus ditentukan dengan tepat terlebih dahulu. Kalau nilai k tidak tepat, maka kesimpulan akhir bisa jauh meleset. Secara spesifik, apabila k lebih kecil dari nilai sebenarnya, maka bisa terjadi *swamping effect*. Sedangkan apabila k lebih besar dari yang seharusnya, *masking effect* yang akan bisa terjadi.

Kendala tersebut diatasi di dalam GESD dengan tidak menuntut nilai k yang tepat. Cukup diketahui batas atasnya saja; batas atas dari banyaknya *outlier*. Misalnya, sebut saja batas atas itu K. Sebagai ilustrasi, jika nilai k (yang sebenarnya tidak diketahui) adalah 5, maka pada teknik Rosner cukup kita ambil batas atas dari k; umpamanya K = 10. Tentang nilai K ini akan diperjelas melalui contoh yang akan disajikan di bab ini.

2. Proses Pengujian Dan Titik Kritis

Misalkan kita telah menentukan batas atas dari nilai k, yakni K. Teknik GESD dilaksanakan sebanyak K langkah pengulangan berturut-turut. Caranya sebagai berikut. Diawali dengan Langkah-1, menguji kehadiran 1 buah *outlier*. Lalu dilanjutkan dengan Langkah-2, menguji 2 *outlier*, …, dan seterusnya dilanjutkan sampai dengan Langkah-K, menguji K buah *outlier*. Jadi, hipotesis yang kita uji adalah:

H0: Tidak ada *outlier* di dalam sekelompok data
H1: Ada paling banyak K buah *outlier* di dalam tumpukan data itu

Statistik pengujinya kita sebut saja $R_{(1)}$, $R_{(2)}$, …, $R_{(K)}$ yang nilainya diberikan pada Langkah-1 sampai dengan Langkah-K di bawah ini.

Langlah-1: $R_{(1)}$ = nilai terbesar di antara n besaran $ABS(X_1-X_{bar1})/s_1$, $ABS(X_2-X_{bar1})/s_1$, …, $ABS(X_n-X_{bar1})/s_1$. Di sini, X_{bar1} dan s_1 adalah mean dan deviasi standar dari semua data orisinal

Langkah-2: $R_{(2)}$ = nilai terbesar di antara n besaran $ABS(X_1-X_{bar2})/s_2$ $ABS(X_2-X_{bar2})/s_2$, …, $ABS(X_n-X_{bar2})/s_2$. Di sini, X_{bar2} dan s_2 adalah mean dan deviasi standar dari semua data tanpa data yang memberikan $R_{(1)}$

Langkah-3: $R_{(3)}$ = nilai terbesar di antara n buah data $ABS(X_1-X_{bar3})/s_3$, $ABS(X_2-X_{bar3})/s_3$, …, $ABS(X_n-X_{bar3})/s_3$. Di sini, X_{bar3} dan s_3 adalah mean dan deviasi standar dari semua data tanpa data yang memberikan $R_{(1)}$ dan $R_{(2)}$

.
.
.....

Langkah-K: $R_{(K)}$ = nilai terbesar di antara n buah data $ABS(X_1-X_{barK})/s_K$, $ABS(X_2-X_{barK})/s_K$, …, $ABS(X_n-X_{barK})/s_K$. Di sini, X_{barK} dan s_K adalah mean dan deviasi standar dari semua data tanpa data yang memberikan $R_{(1)}$, $R_{(2)}$, …, sampai dengan $R_{(K-1)}$

Titik kritis pada setiap langkah kita sebut saja $\lambda_{(1)}$, $\lambda_{(2)}$, …, $\lambda_{(K)}$. Rosner (1983) memberikan nilai aproksimasi untuk titik-titik kritis tersebut dengan menggunakan Formula (1) berikut.

$$\lambda_{(k)} = N_k/D_k \qquad (1)$$

dengan $N_k = (n-k)*q_k$ dan $D_k = SQRT\{(n-k-1+q_k^2)*(n-k+1)\}$ di mana,

1) Nilai k bergerak dari 1, 2, …, sampai dengan K
2) Besaran q_k adalah kuantil ke-100*p dari distribusi-t dengan derajat kebebasan (n-k-1) pada Langkah-k
3) Untuk menghitung q_k, nilai p diberikan oleh Formula (2) sebagai berikut,

$$p = 1 - \alpha/\{2*(n-k+1)\} \qquad (2)$$

4) α adalah tingkat signifikansi yang diinginkan.

Nah, banyaknya *outlier* k sama dengan nilai k terbesar sehingga $R_{(k)} > \lambda_{(k)}$; k = 1, 2, …, K.

Teknik Rosner memiliki keleluasaan (*flexibility*) dalam penggunaannya. Melalui studi simulasi, Rosner (1983) mengingatkan bahwa titik-titik kritis $\lambda_{(k)}$; k = 1, 2, …, K, memberikan keputusan yang sangat akurat untuk $n \geq 25$ dan cukup akurat untuk $n \geq 15$.

Catatan:
Sebenarnya teknik Rosner tidak lain adalah teknik Grubbs yang diterapkan secara sekuensial, namun dengan sedikit perbedaan. Apabila teknik Rosner mampu menangani *masking effect*, tidak demikian halnya dengan teknik Grubbs secara sekuensial. Jika dalam data yang kita analisis ada potensi *masking effect*, penggunaan teknik Grubbs tersebut dapat berakibat hipotesis H0 tidak ditolak padahal seharusnya ditolak. Fenomena ini akan tampak jelas pada contoh di bagian ketiga berikut.

3. Teknik Rosner Dalam Praktik

3.1. Analisis tentang data Rosner

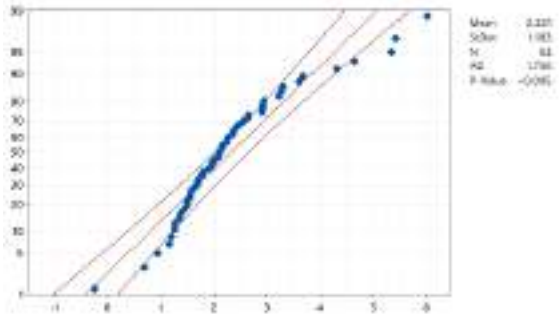
Data acak sebanyak n = 54 pada Tabel 12.1 digunakan dalam Rosner (1983) untuk memperlihatkan keunggulan teknik Rosner dibandingkan dengan penggunaan teknik Grubbs secara sekuensial.

Tabel 12.1. Data acak terurut sebanyak n = 54

No.	Data	No.	Data	No.	Data	No.	Data	No.	Data
1	-0,25	12	1,49	23	1,94	34	2,35	45	3,21
2	0,68	13	1,55	24	1,96	35	2,37	46	3,26
3	0,94	14	1,56	25	1,99	36	2,40	47	3,30
4	1,15	15	1,58	26	2,06	37	2,47	48	3,59
5	1,20	16	1,65	27	2,09	38	2,54	49	3,68
6	1,26	17	1,69	28	2,10	39	2,62	50	4,30
7	1,26	18	1,70	29	2,14	40	2,64	51	4,64
8	1,34	19	1,76	30	2,15	41	2,90	52	5,34
9	1,38	20	1,77	31	2,23	42	2,92	53	5,42
10	1,43	21	1,81	32	2,24	43	2,92	54	6,01
11	1,49	22	1,91	33	2,26	44	2,93		

Sebagai langkah awal, kita buat diagram probabilitas normal. Dengan menggunakan MINITAB diperoleh diagram pada Gambar 12.1 yang menunjukkan bahwa normalitas data patut dipertanyakan. Mengapa menjadi bahan pertanyaan? Sebab, uji kenormalan Anderson-Darling

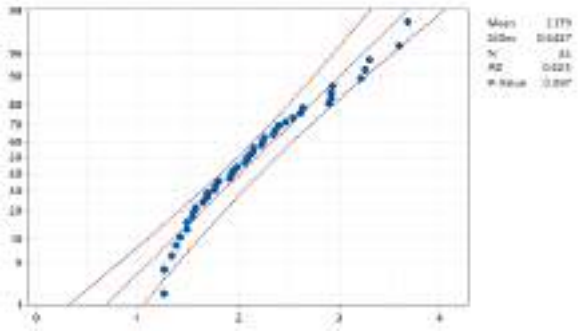
memberikan p-Value < 0,005. Sedangkan, untuk tingkat keberartian 5%, kenormalan data tidak ditolak hanya apabila p-Value > 0,05.



Gambar 12.1. Diagram probabilitas normal untuk n = 54 dengan daerah konfidensi 95%

3.2. Hasil teknik Rosner

Pada Gambar 12.1 tampak ada 5 buah titik data di ekstrim kanan dan 5 buah di ekstrim kiri yang berada di luar atau di batas daerah konfidensi. Apabila kesepuluh data tersebut dikeluarkan, dan untuk 44 buah data sisanya kita buat lagi diagram probabilitas normal, maka kita peroleh diagram pada Gambar 12.2 di bawah ini. Nah, gambar ini memberikan p-Value = 0,097 > 0,05.



Gambar 12.2. Diagram probabilitas normal untuk n = 44 data bagian tengah

Pertanyaannya, apakah memang di dalam data pada Tabel 12.1 di atas ada 10 buah *outlier*? Kita belum tahu. Namun, dari gambar itu cukup beralasan kita mengambil batas atas K = 10. Artinya, di dalam data itu diduga ada sebanyak k buah outlier di mana $k \leq 10$.

Berbekal K = 10, selanjutnya kita ikuti Langkah-1, Langkah-2, …, sampai dengan Langkah-10 seperti yang dilakukan Rosner (1983) tersebut di atas. Hasil selengkapnya yang berupa nilai $R_{(1)}$ sampai dengan $R_{(10)}$ dan nilai $\lambda_{(1)}$ sampai dengan $\lambda_{(10)}$ disajikan pada Tabel 12.2.

Tabel 12.2. Nilai $R_{(1)}$ dan $\lambda_{(1)}$ sampai dengan $R_{(10)}$ dan $\lambda_{(10)}$ untuk α = 5%

k	$n_{(k)}$	$R_{(k)}$	TO	n-k-1	p	q_k	N_k	D_k	$\lambda_{(k)}$
1	54	3.11891	54	52	0.999537	3.513086	186.1936	58.94452	3.15879
2	53	2.94297	53	51	0.999528	3.511063	182.5753	57.93411	3.15143
3	52	3.17942	52	50	0.999519	3.509051	178.9616	56.92362	3.14389
4	51	2.81018	51	49	0.999510	3.507051	175.3526	55.91306	3.13616
5	50	2.81558	1	48	0.999500	3.505068	171.7483	54.90241	3.12825
6	49	2.84817	50	47	0.999490	3.503104	168.1490	53.89170	3.12013
7	48	2.27933	49	46	0.999479	3.501162	164.5546	52.88091	3.11180
8	47	2.31037	48	45	0.999468	3.499247	160.9654	51.87005	3.10324
9	46	2.10158	2	44	0.999457	3.497364	157.3814	50.85913	3.09446
10	45	2.06718	47	43	0.999444	3.495516	153.8027	49.84816	3.08542

Catatan:
Kolom-1 sampai dengan Kolom-10 pada Tabel 12.2 memiliki arti sbb.

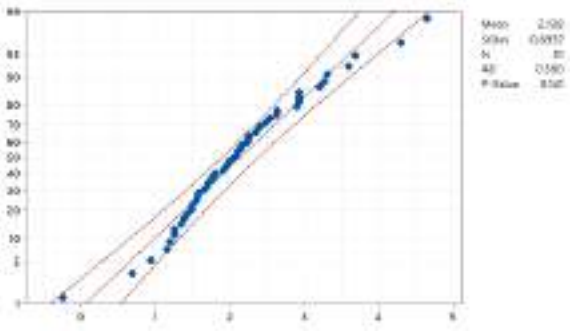
1. Kolom-1: k menyatakan nomor Langkah-k
2. Kolom-2: $n_{(k)}$ adalah ukuran sampel pada Langkah-k
3. Kolom-3: $R_{(k)}$ adalah nilai statistik penguji pada Langkah-k
4. Kolom-4: TO menyatakan nomor data yang menjadi tersangka *outlier*

5. Kolom-5: n-k-1 = 54-k-1 yakni derajat kebebasan pada Langkah-k
6. Kolom-6: p adalah nilai p pada Formula (2)
7. Kolom-7: q_k adalah kuantil ke-100*p dari distribusi-t dengan derajat kebebasan (n-k-1)
8. Kolom-8: N_k adalah nilai pembilang pada Formula (1)
9. Kolom-9: D_k adalah nilai penyebut pada Formula (1)
10. Kolom-10: $\lambda_{(k)}$ adalah nilai titik kritis pada Langkah-k.

Pada proses komputasi di atas jelas terdapat 10 kali pengujian berturut-turut. Dan, pada Langkah-3 kita peroleh nilai $R_{(3)}$ = 3.17942 yang lebih besar dari $\lambda_{(3)}$ = 3.14389. Ini berarti, ada 3 buah outlier di dalam data pada Tabel 12.1 (pada tingkat signifikansi 5%), yakni data nomor 52, 53 dan 54 (tampak di Kolom-4).

Untuk lebih meyakinkan temuan ini, apabila kita buat diagram probabilitas tanpa melibatkan ketiga *outlier* tersebut, maka hasilnya tampak pada Gambar 12.3.

Sangat penting untuk dicatat bahwa Gambar 12.3 mendukung temuan tentang kehadiran 3 *outlier*. Tampak pada gambar itu bahwa, tanpa data nomor 52-54, kenormalan data tidak dapat ditolak pada tingkat dignifikansi 5%. Hal ini ditunjukkan oleh p-Value = 0,141 yang lebih besar dari 5%.



Gambar 12.3. Diagram probabilitas normal untuk semua data tanpa data nomor 52-54

Itulah antara lain kelebihan yang ditawarkan oleh teknik Rosner. Perhatikan baik-baik bahwa p-Value ini lebih bagus ketimbang p-value yang diberikan pada Gambar 12.2 yakni p-Value = 0,097.

3.3. Hasil teknik Grubbs secara sekuensial

Contoh di atas memperlihatkan kehadiran 3 buah *outlier* menurut teknik Rosner pada tingkat signifikansi 5%. Di lain pihak, apabila kita menggunakan teknik Grubbs secara sekuensial, pada sekuen pertama teknik tersebut sudah akan menyatakan TIDAK ada *outlier* (silahkan kerjakan sendiri sebagai latihan).

3.4. Analisis tentang data Grubbs berdasarkan teknik Tietjen-Moore

Sekarang mari kita analisis kembali data Grubbs (1950) yang digunakan Tietjen-Moore (1972) dan telah kita bahas pada Bab 11. Tietjen-Moore menemukan ada dua *outlier* dalam data tersebut (lihat kembali Tabel 10.1 pada Bab 10). Namun, kita hanya menemukan satu buah *outlier* dan bukan dua.

Nah, dua pisau analisis yang berbeda telah memberikan keputusan yang berbeda. Timbullah keragu-raguan; keputusan mana yang lebih bijak? Untuk mengeliminasi keragu-raguan tersebut, kita memerlukan pisau analisis ketiga. Dan, dalam hal ini kita gunakan teknik Rosner untuk dapat menjawab pertanyaan: "Berapa banyak *outlier* di dalam data Grubbs itu?" Berikut ini adalah jawabannya.

Berdasarkan analisis pada Bab 11, maka cukup beralasan kita menduga banyaknya *outlier* k tidak lebih dari 3. Artinya, batas atas banyaknya *outlier* adalah K = 3. Berbekal nilai K = 3, selanjutnya kita laksanakan Langkah-1, Langkah-2, dan Langkah-3 seperti yang telah kita lakukan untuk data Rosner (1983) tersebut di atas. Hasilnya, yang berupa nilai $R_{(1)}$ sampai dengan $R_{(3)}$ dan nilai $\lambda_{(1)}$ sampai dengan $\lambda_{(3)}$, disajikan pada Tabel 12.3.

Tabel 12.3. Nilai $R_{(1)}$ dan $\lambda_{(1)}$ sampai dengan $R_{(3)}$ dan $\lambda_{(3)}$ untuk $\alpha = 5\%$

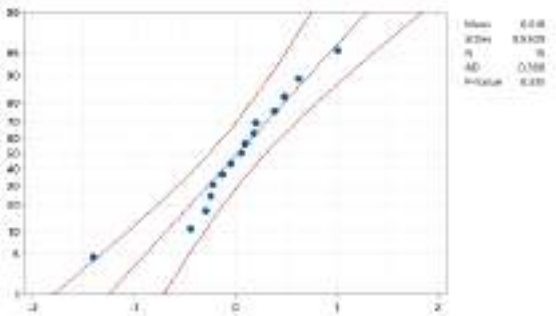
k	$n_{(k)}$	$R_{(k)}$	TO	n-k-1	p	q_k	N_k	D_k	$\lambda_{(k)}$
1	1	2,57374	1	13	0,998333	3,583839	50,17375	19,68905	2,54831
2	15	2,21865	15	12	0,998214	3,611249	46,94624	18,72367	2,50732

3	14	1,80126	14	11	0,998077	3,646204	43,75445	17,77168	2,46203

Tabel ini memperlihatkan kepada kita bahwa pada Langkah-1 kita peroleh nilai $R_{(1)}$ = 2,57374 yang lebih besar dari $\lambda_{(1)}$ = 2,54831. Sedangkan $R_{(2)} < \lambda_{(2)}$ dan $R_{(3)} < \lambda_{(3)}$. Ini berarti, ada 1 buah *outlier* di dalam data itu, yakni data nomor 1. Dengan hasil dari teknik Rosner ini, maka polemik "ada dua *outlier* menurut Tietjen-Moore" dan "ada satu *outlier* menurut analisis penulis di Bab 11" menjadi terang. Dan, teknik Rosner mendukung hasil analisis yang penulis lakukan.

Untuk lebih menajamkan temuan ini, mari kita lihat diagram probabilitas normal pada Gambar 12.4 – Gambar 12.6. Gambar pertama melibatkan kelimabelas data, sedangkan gambar kedua hanya melibatkan 14 buah data tanpa data nomor 1 yang berupa *outlier*. Selanjutnya, pada Gambar 12.6 hanya 13 buah data yang dilibatkan; tanpa data nomor 1 dan nomor 15. Hasilnya sangat menarik untuk dicatat.

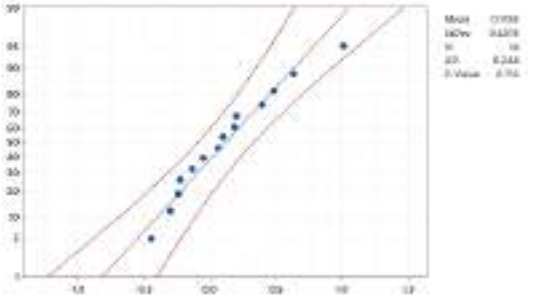
1. Kenormalan data dapat diterima walaupun di dalam data itu ada *outlier*. Fenomena ini ditunjukkan oleh Gambar 12.4 dengan p-Value = 0,381 yang menunjukkan tingkat kenormalannya.



Gambar 12.4. Diagram probabilitas normal untuk semua data n = 15

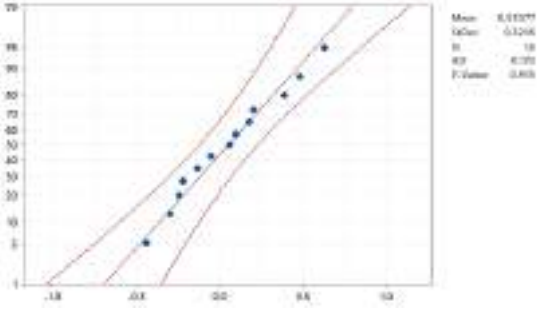
2. Tatkala data *outlier* nomor 1 tidak dilibatkan, maka kenormalan data langsung meningkat banyak dengan p-

Value = 0,712 seperti tampak pada Gambar 12.5. Sekali lagi, kemampuan menangani *outlier* sangat menentukan keputusan berdasarkan analisis statistik.



Gambar 12.5. Diagram probabilitas normal untuk semua data tanpa data nomor 1

3. Selanjutnya, apabila kita mengikuti kaidah Tietjen-Moore yang menganggap data nomor 1 dan 15 sebagai *outlier*, lalu memisahkan kedua data ini dari kelompoknya, kenormalan data meningkat lagi menjadi p-Value = 0,913 seperti tertera pada Gambar 12.6. Namun, sesuai dengan analisis penulis di Bab 11 dan hasil teknik Rosner, perbuatan menganggap kedua data tersebut sebagai *outlier* adalah perbuatan yang berlebihan (*superfluous*).



Gambar 12.6. Diagram probabilitas normal untuk semua data tanpa data nomor 1 dan 15

4. Catatan Penutup

Dari uraian dan contoh di atas, tampak kegunaan lain dari teknik GESD. Ia dapat menjawab pertanyaan: "Berapa banyak data *outlier* k yang ada dalam sekumpulan data?" Dengan menggunakan jawaban tentang nilai k ini, teknik Tietjen-Moore selanjutnya dapat diterapkan untuk lebih menguatkan keputusan tentang hasil pengujian *outlier*. Di samping itu, untuk menghindari potensi terjadinya *masking effect* dan *swamping effect*, penggunaan teknik Grubbs secara sekuensial tidak dianjurkan. Sebagai gantinya, gunakanlah teknik Rosner.

Itulah 5 buah teknik yang kami sarankan untuk dijadikan bahan pertimbangan utama tatkala hendak menguji kehadiran *outlier* di dalam sekelompok data univariat. Sekali lagi, kelima teknik itu adalah:

1. Teknik Grubbs (ESD) yang diperkenalkan oleh Grubbs (1950)
2. Teknik IESD yang tidak lain adalah teknik ESD dengan titik kritis yang eksak (Djauhari, 2001a, 2001b, 2003)
3. Teknik Dixon (r_{10}, r_{11}, r_{12}, r_{20}, r_{21} dan r_{22}) yang dikemukakan oleh Dixon (1950)
4. Teknik Tietjen-Moore (L_{ka}, L_{ki} & E_k) karya Tietjen-Moore (1972)
5. Teknik Generalized ESD (GESD) karya Rosner (1983).

Walaupun demikian, masih ada dua pertanyaan besar yang perlu dijawab yakni:

1. Kelima teknik yang kita bahas mulai Bab 8 sampai dengan Bab 12 ini dimaksudkan untuk ukuran sampel n kecil hingga n cukup besar. Akan tetapi tidak untuk n yang terlalu besar. Bagaimana kalau n sangat besar? Jawaban pertanyaan ini, untuk kasus di mana data berasal dari distribusi normal, akan dikemukakan pada Bab 13.
2. Bagaimana kalau data yang kita analisis adalah data multivariat? Nah, pertanyaan ini kita simpan untuk dibahas pada buku jilid kedua yang khusus membahas pengujian data *outlier* multivariat.

BAB 13. UJI KESAHIHAN: TEKNIK FAST MINIMUM VARIANCE

"Anyone who has never made a mistake has never tried anything new."

Albert Einstein

Pembersihan data dari kehadiran *outlier* adalah aktivitas yang mutlak harus dilaksanakan sebelum memulai setiap analisis data dan setiap analisis statistik. Seperti telah dikemukakan di bab-bab sebelumnya, aktivitas ini terdiri atas tiga tahap berikut.

1. Tahap penyelidikan; mengidentifikasi calon tersangka *outlier*
2. Tahap penyidikan; mendeteksi apakah calon tersangka patut dijadikan tersangka *outlier*
3. Tahap pengadilan; menguji apakah tersangka "bersalah" secara sahih/signifikan sebagai *outlier*.

Dua teknik untuk menyelidiki calon tersangka telah kita bahas, yakni teknik berbasis data terurut dan teknik grafik. Lalu, untuk menyidik apakah calon tersangka patut dijadikan tersangka *outlier*, juga telah kita kemukakan dua teknik. Pertama, teknik Tukey dan kedua teknik Iglewicz-Hoaglin.

Sedangkan untuk menguji kesahihan, 5 teknik yang sangat populer telah kita diskusikan secara rinci di lima bab sebelumnya yakni Bab 8 sampai dengan Bab 12. Kelima teknik itu adalah,

1. Teknik ESD (Grubbs, 1950)
2. Teknik IESD (Djauhari, 2001a, 2001b, 2003)
3. Teknik Dixon (1950)
4. Teknik Tietjen-Moore (1972)
5. Teknik GESD (Rosner, 1983)

Di antara kelima teknik ini, hanya dua teknik yang bersifat tangguh yakni teknik Dixon dan teknik Tietjen-Moore. Sayangnya, kedua teknik ini hanya cocok untuk n yang kecil (teknik Dixon) dan n yang

tidak terlalu besar (teknik Tietjen-Moore). Bagaimana kalau n besar? Nah, untuk tujuan itulah pada bab ini kita akan bicara tentang teknik yang tangguh dan sejauh pemahaman kami belum pernah ada sebelumnya (*unprecedented*) secara khusus, dan sekaligus dapat digunakan untuk n yang sangat besar atau bahkan yang berupa *big data*. Teknik ini merupakan kontribusi kami para penulis kepada pengembangan dan kemajuan ilmu statistik.

Teknik yang akan kita bahas pada bab ini, kami namakan teknik FMV (*Fast Minimum Variance*). Teknik ini adalah hal khusus dari teknik FMCD (*Fast Minimum Covariance Determinant*) dan dari teknik MVV (*Minimum Vector Variance*). Teknik FMCD adalah karya Rousseeuw dan van Driessen (1999) yang sangat populer. Sedangkan teknik MVV diperkenalkan oleh kami sendiri di dalam Herwindiati, Djauhari dan Mashuri (2007) untuk mengatasi keterbatasan dari teknik FMCD. Dengan demikian, dalam praktik, FMCD dan MVV saling komplementer satu sama lain. Penggunaan kedua teknik ini secara bersamaan akan memberikan efek yang saling memperkuat.

Bila MVV dan FMCD dimaksudkan untuk membersihkan data kompleks (multivariat atau multidimensi artinya ada banyak variabel statistik yang saling berkorelasi), FMV adalah hal khusus dari MVV dan FMCD yakni untuk membersihkan data unidimensi (univariat artinya hanya ada 1 variabel statistik yang dikaji). Oleh karena itu, maka semua sifat yang dimiliki FMV adalah turunan dari sifat-sifat yang dimiliki baik oleh MVV maupun oleh FMCD.

Sebagai statistik tangguh, FMCD dan MVV memiliki sifat yang ideal. Inilah keunggulan FMCD dan juga MVV sehingga penggunaannya sangat populer dan dapat ditemukan dalam spektrum bidang ilmu yang amat luas. Bentuk awal FMCD tatkala diperkenalkan pertama kali oleh Rousseeuw (1985) masih sangat primitif, yakni berbentuk MCD (*Minimum Covariance Determinant*) tanpa kata "*Fast*" di depannya. Empatbelas tahun kemudian, barulah Rousseeuw dan van Driessen (1999) memperkenalkan FMCD sebagai pengembangan dari MCD dengan proses komputasi yang jauh lebih cepat. Pengembangan ini dimungkinkan berkat dalil/teorema matematika yang mereka perkenalkan tentang proses mengkonsentrasikan data (*data concentration process*). Sejak saat itu, FMCD terus berkembang. Namun, hingga saat ini pengembangan yang dilakukan

para ahli terfokus pada upaya mempercepat lagi proses komputasi terutama tatkala berhadapan dengan *big data*.

Di samping keunggulannya yang mengesankan, FMCD memiliki kelemahan atau keterbatasan yang bisa sangat mengganggu, yakni adanya syarat bahwa "matriks kovariansi harus non-singular." Syarat ini muncul secara natural sebab FMCD didasarkan kepada CD (*covariance determinant*) atau determinan matriks kovariansi sebagai ukuran dispersi multivariat. Sedangkan determinan sebuah matriks akan ada nilainya (bukan nol), jika matriks tersebut non-singular. Jadi, begitu matriks kovariansi tidak non-singular, maka teknik FMCD yang tangguh itu tidak dapat digunakan. Untuk mengatasi keterbatasan ini, Djauhari (2007) memperkenalkan VV (*vector Variance*) sebagai ukuran disperi multiviat selain CD. Tidak lama kemudian, pada tahun itu juga Herwindiati, Djauhari dan Mashuri (2007) memperkenalkan teknik MVV yang komplementer dengan teknik FMCD.

Selain proses komputasinya lebih cepat, patut dicatat bahwa tatkala matriks kovariansi bersifat non-singular, hasil yang diberikan oleh MVV dan yang diberikan oleh FMCD, kedua-duanya saling melengkapi. Oleh karena itu, penggunaan FMCD dan MVV secara bersama-sama adalah sikap yang bijak. Khusus dalam kasus data univariat, FMCD dan MVV mendefinisikan FMV yang sama.

Teori tentang FMCD dan juga MVV amat sangat jelimet. Oleh karena itu, teori tersebut kami berikan secara mendetil hanya kepada mahasiswa program Doktoral atau Post-Doktoral bidang ilmu statistika dengan latar belakang matematika yang kuat. Untuk keperluan praktis, termasuk bagi para praktisi sains sosial, di sini kami hanya akan memberikan algoritma teknik FMV. Bagi yang tertarik mempelajari teorinya, silahkan lihat referensi tersebut di atas yang tertera di daftar referensi di akhir buku ini. Atau silahkan kontak kami.

Algoritma FMV terdiri atas dua tahap, yakni 1) tahap mengkonsentrasikan data (*data concentration*) dan mengurutkan data mulai dari pusat data (rata-rata) menuju ke arah luar yakni ke arah data ekstrim (*centre-outward ordering*), dan 2) tahap menentukan titik kritis.

1. Tahap Mengkonsentraskan & Mengurutkan Data

Tahap ini terdiri atas sebelas langkah berikut.

1. Di antara n buah data acak X_1, X_2, …, X_n yang berasal dari distribusi normal, pilih h buah data di mana h = bilangan bulat terbesar yang lebih kecil dari atau sama dengan 3n/4 (75% data).
2. Hitung rata-rata (sebut saja, X_{barO}) dan deviasi standar (S_O) dari h buah data terpilih.
3. Hitung D_1, D_2, …, D_n di mana,
   1) $D_1 = ((X_1 - X_{barO})/S_O)^2$
   2) $D_2 = ((X_2 - X_{barO})/S_O)^2$
   .
   .
   .
   n) $D_n = ((X_n - X_{barO})/S_O)^2$
4. Urutkan nilai D_1, D_2, …, D_n dari yang terkecil sampai dengan yang terbesar. Hasil pengurutan itu kita sebut $D_{(1)}$, $D_{(2)}$, …, $D_{(n)}$.
5. Ambil $D_{(1)}$, $D_{(2)}$, …, $D_{(h)}$ untuk proses selanjutnya. Lalu, data yang memberikan $D_{(1)}$, $D_{(2)}$, …, $D_{(h)}$ kita sebut saja Y_1, Y_2, …, Y_h.
6. Hitung rata-rata (X_{barN}) dan deviasi standar (S_N) dari Y_1, Y_2, …, Y_h.
7. Jika $X_{barN} = X_{barO}$ dan $S_N = S_O$, teruskan ke Langkah 9. Jika X_{barN} tidak sama dengan X_{barO} dan/atau S_N tidak sama dengan S_O, teruskan ke Langkah 8.
8. Nilai X_{barO} diganti dengan nilai X_{barN} dan nilai S_O diganti dengan nilai S_N. Lalu, kembali ke Langkah 3.
9. Proses mengkonsentrasikan data selesai.
10. Data yang memberikan $D_{(1)}$, $D_{(2)}$, …, $D_{(n)}$ kita sebut saja Y_1, Y_2, …, Y_n.
11. Maka Y_1, Y_2, …, Y_n adalah data X_1, X_2, …, X_n yang sudah terkonsentrasi di sekitar rata-rata yang tangguh (*robust mean sample*) dan terurut dari yang terkecil sampai dengan terbesar dengan skala urutan diberikan oleh $D_{(1)}$, $D_{(2)}$, …, $D_{(n)}$.

Contoh
Pada Tabel 13.1, kolom kedua, diberikan n = 15 buah data acak yang berasal dari distribusi normal. Data ini telah kita analisis pada Bab 8 dengan teknik Grubbs (ESD). Di sini data tersebut akan digunakan untuk memberi gambaran tentang cara mengimplementasikan Langkah 1 sampai dengan Langkah 11 di atas.

Karena n = 15, maka h = 11 (yakni bilangan bulat terbesar yang lebih kecil dari atau sama dengan 3*15/4 = 11,25).

Tabel 13.1. Limabelas buah data acak

No.	Data X1-X15
1	165
2	188
3	194
4	197
5	200
6	202
7	205
8	210
9	214
10	215
11	227
12	231
13	239
14	249
15	297

2. Proses Iterasi
2.1. Iterasi Ke-1
Iterasi ini memberikan hasil yang tertera pada Tabel 13.2 dengan rata-rata (MEAN) h buah data = 201,55 dan deviasi standarnya (SD) = 16,33. Setelah nilai rata-rata dari h buah data dan deviasi standarnya diperoleh, kita lanjutkan dengan;

1. Menghitung nilai D_1 sampai dengan D_{15}. Hasilnya ada pada kolom ke-3, Tabel 13.2
2. Mengurutkan nilai D_1 sampai dengan D_{15} dari yang terkecil hingga terbesar. Hasilnya disajikan pada kolom ke-4
3. Pengurutan tersebut memberikan urutan nomor data dari yang paling dekat ke MEAN hingga yang paling jauh. Hasilnya ada pada kolom ke-5
4. Data yang sudah diurutkan tersebut disimpan di kolom terakhir (Ini tidak lain adalah data X_1 hingga X_{15} yang diurutkan sesuai urutan pada Butir 3 di atas)

Karena urutan h buah data pada kolom ke-1 tidak sama dengan urutan h buah data pada kolom ke-5, maka proses pengonsentrasian data diteruskan ke iterasi ke-2.

Tabel 13.2. Hasil iterasi ke-1 pengkonsentrasian data

No. h data	h data X	D	D terurut	No. Data Y	Data Y
1	165	5,00827461	0,00077478	6	202
2	188	0,68803188	0,00895641	5	200
3	194	0,21349721	0,04475105	7	205
4	197	0,07747758	0,07747758	4	197
5	200	0,00895641	0,21349721	3	194
6	202	0,00077478	0,26804143	8	210
7	205	0,04475105	0,58167067	9	214
8	210	0,26804143	0,67882755	10	215
9	214	0,58167067	0,68803188	2	188
10	215	0,67882755	2,42969685	11	227
11	227	2,42969685	3,25331449	12	231
		3,25331449	5,00827461	1	165
		5,26054160	5,26054160	13	239
		8,44456014	8,44456014	14	249
		34,16761189	34,16761189	15	297
MEAN	201.55				
SD	16.33				

2.2. Iterasi Ke-2

Pada iterasi ke-2, kita gunakan h buah data pertama hasil pengurutan pada iterasi ke-1. Nomor h buah data pertama yang ada di kolom ke-5 dan nilai h buah data pertama yang ada di kolom ke-6 pada Tabel 13.2 kita simpan di kolom ke-1 dan ke-2 pada Tabel 13.3. Data ini memberikan MEAN = 207,55 dan SD = 13,43.

Selanjutnya,

1. Kita hitung nilai D_1 sampai dengan D_{15}. Hasilnya ada pada kolom ke-3, Tabel 13.3
2. Lalu nilai D_1 sampai dengan D_{15} ini diurutkan dari yang terkecil hingga terbesar. Hasilnya disajikan pada kolom ke-4
3. Pengurutan tersebut memberikan urutan nomor data dari yang paling dekat ke MEAN hingga yang paling jauh. Hasilnya ada pada kolom ke-5
4. Data yang sudah diurutkan sesuai nomor pada kolom ke-5 disimpan di kolom terakhir (Ini tidak lain adalah data X_1 hingga X_{15} yang diurutkan pada iterasi ke-2)

Tabel 13.3. Hasil iterasi ke-2 pengkonsentrasian data

No. h data	h data X	D	D terurut	No. Data Y	Data Y
6	202	0,17058635	0,03342044	8	210
5	200	0,31582084	0,03594187	7	205
7	205	0,03594187	0,17058635	6	202
4	197	0,61687984	0,23110072	9	214
3	194	1,01778756	0,30825654	10	215
8	210	0,03342044	0,31582084	5	200
9	214	0,23110072	0,61687984	4	197
10	215	0,30825654	1,01778756	3	194
2	188	2,11914913	2,09948196	11	227
11	227	2,09948196	2,11914913	2	188
12	231	3,05157475	3,05157475	12	231
		10,04098473	5,48828680	13	239
		5,48828680	9,53266401	14	249

		9,53266401	10,04098473	1	165
		44,3889423 7	44,38894237	15	297
MEAN	207.55				
SD	13.43				

Karena iterasi ke-2 ini memberikan nilai SD yang berbeda dengan nilai SD pada iterasi ke-1 (tepatnya lebih kecil sesuai teori yang melandasinya), maka proses pengkonsentrasian data dilanjutkan lagi ke iterasi ke-3.

2.3. Iterasi Ke-3

Pada iterasi ini, kita gunakan h buah data pertama hasil pengurutan pada iterasi ke-2. Nomor h buah data pertama yang ada di kolom ke-5 dan nilai h buah data pertama yang ada di kolom ke-6 pada Tabel 13.3 kita simpan di kolom ke-1 dan ke-2 pada Tabel 13.4. Data ini memberikan MEAN = 207,55 dan SD = 13,43.

Karena nilai MEAN dan SD pada iterasi ini sama dengan nilai MEAN dan SD pada iterasi sebelumnya, maka proses pengkonsentrasian data dihentikan. Dengan kata lain, Tahap 1 telah selesai dilaksanakan.

Walaupun SD pada iterasi ke-3 sama dengan SD pada iterasi ke-2, perlu dicatat adanya perbedaan urutan h buah data yang sudah terkonsentrasi antara kedua iterasi tersebut. Urutan pada iterasi ke-2 adalah 6, 5, 7, 4, 3, 8, 9, 10, 2, 11, dan 12. Sedangkan pada iterasi ke-3 adalah 8, 7, 6, 9, 10, 5, 4, 3, 11, 2, dan 12.

Tabel 13.4. Hasil iterasi ke-3 pengkonsentrasian data

No. h data	h data X	D	D Terurut	No. Data Y	Data Y
8	210	0,03342044	0,03342044	8	210
7	205	0,03594187	0,03594187	7	205
6	202	0,17058635	0,17058635	6	202

9	214	0,23110072	0,23110072	9	214
10	215	0,30825654	0,30825654	10	215
5	200	0,31582084	0,31582084	5	200
4	197	0,61687984	0,61687984	4	197
3	194	1,01778756	1,01778756	3	194
11	227	2,09948196	2,09948196	11	227
2	188	2,11914913	2,11914913	2	188
12	231	3,05157475	3,05157475	12	231
		5,48828680	5,48828680	13	239
		9,53266401	9,53266401	14	249
		10,04098473	10,04098473	1	165
		44,38894237	44,38894237	15	297

MEAN	207.55
SD	13.43

Pada tabel ini tampak bahwa:

1. Urutan nomor data yang sudah terkonsentrasi adalah 8, 7, 6, 9, 10, 5, 4, 3, 11, 2, 12, 13, 14, 1, dan 15.
2. Berbeda dengan tabel-tabel sebelumnya, pada Tabel 13.4 ini, kolom ke-3 sama dengan kolom ke-4. Ini menandakan bahwa data sudah terkonsentrasi dan terurut berdasarkan jauhnya dari MEAN dalam satuan SD.

Nah, dalam literatur, MEAN = 207,55 dan SD = 13,43 yang diperoleh pada iterasi terakhir dinamakan *robust sample mean* dan *robust standard deviation*. Selanjutnya, data terurut pada kolom terakhir dinamakan *center-outward ordered data*.

Dengan selesainya proses pengonsentrasian data dan pengurutan data dari pusat keluar, pada kolom terakhir Tabel 13.4 disajikan data yang sudah terkonsentrasi dan terurut berdasarkan letaknya atau jauhnya dari rata-rata yang tangguh sebagai titik pusat (*central*). Urutannya diberikan pada kolom ke-5 dan skala urutannya pada kolom ke-4. Jadi, kita baca, data No. 8 (210) adalah yang paling dekat dari rata-

rata yang tangguh (= MEAN = 207,55), yakni sejauh SQRT(0.033420437)*SD dengan SD = 13,43. Sedangkan data No. 15 (297) adalah yang paling jauh, yakni sejauh SQRT(44.38894237)*SD.

Contoh ini memperlihatkan bagaimana tahap pengonsentrasian data dan pengurutan data bekerja. Setelah tahap ini selesai dilaksanakan dan telah diperoleh data tentang D_1, D_2, …, D_n, maka tugas selanjutnya adalah:

1. Menguji hipotesis H0: "Tidak ada *outlier* dalam data" dengan H1: "Ada *outlier* dalam data"
2. Mengenalpasti semua *outlier* yang ada, jika H0 ditolak

Di dalam literatur, D_1, D_2, …, D_n disebut kuadrat jarak Mahalanobis yang tangguh (*robust Mahalanobis Squared Distance*, disingkat RMSD). Nah, kedua tugas tersebut baru akan dapat kita kerjakan setelah distribusi RMSD berhasil diidentifikasi dan kemudian titik kritis statistik penguji FMV dapat dihitung. Masalah ini akan menjadi topik bahasan pada Tahap 2 (Penentuan titik kritis). Namun, untuk sementara, silahkan tingkatkan kemahiran dalam melakukan pengonsentrasian dan pengurutan data dengan menggunakan data yang dimiliki pembaca.

Apabila pembaca menghadapi masalah, jangan segan untuk mendiskusikannya dengan kami. Jangan lupa, data yang bersih dan analisis yang tajam akan menjamin hasil analisis data dan analisis statistik yang akurat dan handal (*reliable*). Dan, pada akhirnya tentu akan meningkatkan reputasi pengguna buku ini.

3. Tahap Penentuan Titik Kritis

Kembali kepada pertanyaan: "Adakah *outlier* di antara kelompok data yang kita analisis?" Lalu, "Jika ada, sebutkan yang mana?" Kedua pertanyaan ini membawa kita kepada masalah tentang,

1. Distribusi statistik dari RMSD; D_1, D_2, …, D_n.
2. Titik kritis yang bersesuaian dengan distribusi RMSD.

Dalam literatur statistika modern, usia distribusi RMSD tergolong sangat baru. Ia baru berusia 16 tahun. Tatkala FMCD diperkenalkan

pertama kali oleh Rousseeuw and van Driessen (1999), distribusi RMSD masih dinyatakan dalam bentuk distribusi limitnya, yakni distribusi Chi-kuadrat dengan derajat kebebasan v, ditulis $\chi^2(v)$, di mana v tidak lain adalah banyaknya variabel yang terlibat dalam analisis data multivariat. Dalam kasus univariat yang menjadi topik bahasan dalam buku ini, v = 1. Artinya, dalam hal ini, distribusi limit dari RMSD adalah $\chi^2(1)$.

Karena digunakan distribusi limit, tentu titik kritisnya akan melenceng dari yang sebenarnya khususnya tatkala ukuran sampel n kecil. Untuk menghindari adanya kesalahan ini, pada tahun 2005, dua sekawan Hardin dan Rocke berhasil mengidentifikasi distribusi RMSD secara eksak. Dalam artikel Hardin dan Rocke (2005), mereka membuktikan bahwa RMSD berdistribusi sebagai berikut.

$$\text{RMSD berdistribusi } (vm)F(v, m - v + 1)/\{c(m - v + 1)\}$$

di mana m adalah derajat kebebasan dari matriks kovariansi sampel yang tangguh dan c adalah konstanta pembagi yang membuat matriks tersebut menjadi penaksir tak bias untuk matriks kovariansi populasi. Sedangkan F(v, m – v +1) menyatakan distribusi F (Fisher) dengan derajat kebebasan pembilang v dan derajat kebebasan penyebut (m – v +1).

Apabila dihitung sejak Rousseeuw mendefinisikan RMSD sebagai solusi MCD tahun 1985, yang kemudian algoritmanya di diperbaharui tahun 1999 bersama van Driessen dalam bentuk FMCD, maka peradaban manusia memerlukan waktu 20 tahun sampai diperolehnya formula di atas.

Dalam kasus yang sedang kita bahas di mana v = 1 (univariat), distribusi RMSD di atas dapat disederhanakan menjadi,

$$\text{RMSD berdistribusi } F(1, m)/c. \qquad (1)$$

Distribusi ini, untuk ukuran sampel yang cukup besar (secara umum n > 1.000), dapat didekati dengan menggunakan distribusi limitnya. Pendekatan ini cukup bagus dalam arti “beda antara nilai eksak dengan nilai pendekatan berada pada angka kedua di belakang koma desimal.” Dengan demikian, bisa kita gunakan RMSD berdistribusi

$\chi^2(1)$. Sedangkan untuk n yang kecil (secara umum n < 1.000), nilai konstanta c = 1 dan derajat kebebasan m diberikan oleh Hardin dan Rocke (2005) sebagai berikut,

$$m = M \cdot e^{(0{,}725 - 0{,}00663 - 0{,}0780 \ln(n))} \quad (2)$$

di mana e adalah bilangan yang memenuhi ln(e) = 1 atau e = 2,7828 dan M dihitung dengan menggunakan 10 langkah di bawah ini.

1. $c_\alpha = (1 - \alpha)/P(\chi^2(3) \leq q_\alpha)$
2. $c_2 = -P(\chi^2(3) \leq q_\alpha)/2$
3. $c_3 = -P(\chi^2(5) \leq q_\alpha)/2$
4. $c_4 = 3c_3$
5. $b_1 = c_\alpha (c_3 - c_4)/(1 - \alpha)$
6. $b_2 = 0{,}5 + \{c_\alpha/(1 - \alpha)\}\{c_3 - q_\alpha [c_2 + (1 - \alpha)/2]\}$
7. $v_1 = (1 - \alpha)b_1^2(\alpha(c_\alpha q_\alpha - 1)^2 - 1) - 2c_3c_\alpha^2(3(b_1 - b_2)^2 + 3b_2(2b_1 - b_2))$
8. $v_2 = n(b_1(b_1 - b_2)(1 - \alpha))^2c_\alpha^2$
9. $v = v_1/v_2$
10. $M = 2/(c_\alpha^2 v)$

Berdasarkan formula (1) dan (2), maka titik kritis C untuk pengujian H0 misalnya dengan tingkat signifikansi α = 5% dan m = 7 diberikan oleh MS Excel melalui perintah berikut: "= F.INV(1-0.05,1,7)". Untuk nilai α dan m yang lain, misalnya α = 1% dan m = 10, tentu perintahnya adalah "= F.INV(1-0.01,1,10)".

Catatan:
Proses komputasi nilai m pada (2) sangat kompleks dan jelimet. Dituntut ekstra hati-hati untuk melakukannya. Apalagi untuk menurunkannya secara matematis. Namun, yang terakhir ini biarlah para statistisi matematikal (mathematical statisticians) yang menanganinya. Satu hal yang perlu dicatat, bila m yang diberikan oleh (2) bukan berupa bilangan bulat, maka m perlu dibulatkan ke bilangan bulat terdekat terlebih dahulu.

Untuk memudahkan kerja para pembaca, pada Tabel 13.5 berikut kami sajikan nilai titik kritis C untuk berbagai nilai n dengan tingkat signifikansi 1%, 2,5%, 5% dan 10%.

Tabel 13.5. Titik kritis RMSD

n	Tingkat signifikasi			
	1%	2,5%	5%	10%
10	8,39974	6,04201	4,45132	3,02623
11	8,18495	5,92163	4,38075	2,98990
12	8,01660	5,82665	4,32479	2,96096
13	7,94539	5,78630	4,30095	2,94858
14	7,82287	5,71664	4,25968	2,92712
15	7,72125	5,65862	4,22520	2,90913
16	7,63562	5,60956	4,19597	2,89385
17	7,59905	5,58855	4,18342	2,88727
18	7,56248	5,56753	4,17088	2,88069
19	7,49928	5,53113	4,14910	2,86926
20	7,44414	5,49929	4,13002	2,85922
25	7,31410	5,42394	4,08475	2,83535
30	7,19422	5,35413	4,04265	2,81308
35	7,11029	5,30506	4,01297	2,79733
40	7,04825	5,26868	3,99092	2,78560
45	7,01140	5,24703	3,97778	2,77860
50	6,97140	5,22348	3,96347	2,77098
60	6,91242	5,18870	3,94230	2,75967
70	6,87782	5,16826	3,92984	2,75301
80	6,84727	5,15019	3,91882	2,74711
90	6,82640	5,13782	3,91127	2,74306
100	6,80689	5,12626	3,90420	2,73927
125	6,77541	5,10758	3,89278	2,73315
150	6,75311	5,09433	3,88467	2,72879
175	6,73692	5,08470	3,87877	2,72563
200	6,72495	5,07758	3,87441	2,72328
250	6,70826	5,06764	3,86832	2,72001

300	6,69664	5,06072	3,86407	2,71772
350	6,68853	5,05589	3,86111	2,71613
400	6,68222	5,05213	3,85880	2,71489
450	6,67738	5,04924	3,85703	2,71393
500	6,67337	5,04685	3,85557	2,71314
600	6,66740	5,04329	3,85338	2,71197
700	6,66306	5,04070	3,85179	2,71111
800	6,65980	5,03875	3,85059	2,71046
900	6,65723	5,03722	3,84965	2,70996
1000	6.65516	5,03598	3,84889	2,70955
∞	6.63490	5,02389	3,84146	2,70554

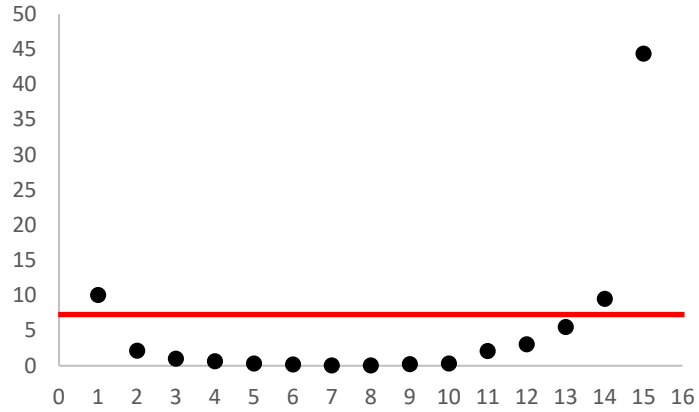
Pada tabel ini tampak bahwa untuk tingkat signifikansi 1%, beda antara titik kritis eksak dengan titik kritis pendekatan ($n = \infty$) terletak pada digit kedua di belakang koma tatkala $n = 300$ atau lebih. Namun, untuk tingkat signifikansi 2,5%, 5% dan 10%, beda seperti tersebut dicapai tatkala n paling kecil sama dengan 150, 125 dan 35. Dalam praktik, tatkala menggunakan RMSD untuk menguji kehadiran *outlier* dalam sekelompok data, Rousseeuw dan van Driessen (1999) menyarankan penggunaan tingkat signifikansi 2,5%.

Contoh

Perhatikan kembali data pada Tabel 13.1, kolom kedua, yang berasal dari distribusi normal dengan $n = 15$. Pada contoh pertama telah kita peroleh nilai RMSD $D_1, D_2, \ldots, D_{10}$ seperti tampak pada Tabel 13.4, kolom ketiga. Nilai RMSD terbesar adalah 44.38894237 yang diberikan oleh data No. 15.

Selanjutnya, bila kita gunakan tingkat signifikansi 2,5% seperti saran Rousseeuw dan van Driessen (1999), untuk $n = 15$, Tabel 13.5 memberikan titik kritis $C = 5{,}65862$. Dengan demikian, tampak jelas bahwa nilai RMSD terbesar lebih besar dari titik kritis C. Dengan kata lain, H0: "Tidak ada *outlier*" ditolak.

Pertanyaannya, “ada berapa *outlier*? dan “tunjukkan data yang menjadi *outlier*!” Secara visual dengan menggunakan diagram deretan data, pada Gambar 13.1 tampak ada 3 data calon tersangka *outlier* yakni data nomor 1, 14, dan 15 yang memiliki RMSD berturut-turut 10,04098 dan 9,53266 dan 44,38894.



Gambar 13.1. Diagram deretan data RMSD

Pada gambar ini, sumbu horizontal menyatakan nomor data sedangkan sumbu vertikal adalah nilai RMSD. Lalu, garis berwarna merah menyatakan nilai titik kritis C = 5,65862. Tampak jelas bahwa data nomor 1, 14, dan 15 adalah *outlier*.

Hasil ini berbeda dengan hasil yang diberikan teknik ESD maupun teknik IESD. Apabila menurut teknik IESD ada 2 *outlier* yakni data nomor 1 dan 15, maka menurut FMV ada 3 yakni 1, 14 dan 15. Keputusan yang mana yang lebih bisa diterima? Tanpa dibekali dengan informasi tambahan tentang data yang kita analisis, tentu keputusan yang diberikan FMV lebih disukai kalau saja n besar. Namun, n = 15 tentu belum cukup untuk dikatakan besar. Oleh karena itu, keputusan yang bijak adalah yang diberikan oleh Teknik IESD.

Dengan berakhirnya bab ini, maka cukup lengkaplah senjata para pembaca dalam upaya membersihkan data dari kehadiran *outlier*. Tepatnya, ada 6 buah senjata yakni:

1. Teknik Grubbs (ESD) yang diperkenalkan oleh Grubbs (1950)
2. Teknik IESD yang tidak lain adalah uji ESD dengan titik kritis yang eksak (Djauhari, 2001a, 2001b, 2003)
3. Teknik Dixon (r_{10}, r_{11}, r_{12}, r_{20}, r_{21} dan r_{22}) yang dikemukakan oleh Dixon (1950)
4. Teknik Tietjen-Moore (L_{ka}, L_{ki} & E_k) karya Tietjen-Moore (1972)
5. Teknik Generalized ESD (GESD) karya Rosner (1983).
6. Teknik FMV karya tulen kami, versi univariat dari FMCD karya Rousseeuw dan van Driessen (1999), dan dari MVV karya Herwindiati, Djauhari dan Mashuri (2007).

Selamat datang di komunitas pencinta data bermutu!

BAB 14. EPILOG: BAHAN TERAWANGAN

"*If all you have is a hammer, everything looks like a nail*."

Abraham Maslow

Langkah kanan dalam analisis data dan analisis statistik amat sangat menentukan perjalanan berikutnya. Langkah itu ibarat kunang-kunang di kegelapan malam yang menerangi jalan yang akan ditempuh selanjutnya. Pengalaman kami dalam memberikan konsultasi kepada para pengguna ilmu statistika dari berbagai kalangan sejak 1982, baik nasional maupun internasional, menunjukkan bahwa para pengguna ilmu statistika umumnya terjebak dalam situasi yang saya gambarkan lewat ungkapan Abraham Maslow berikut: "*If all you have is a hammer, everything looks like a nail*" yang tertulis di atas. Maksudnya seperti begini. Kalau metode analisis statistik yang diketahui seseorang adalah metode analisis regresi, maka orang tersebut akan cenderung berusaha membangun pola pikir (*mindset*) dan bahkan jalan pikiran (*mind-map*) agar dapat menggunakan metode itu pada data yang dimilikinya.

Tidak boleh begitu! Sekali-kali janganlah begitu. Hormatilah data. Data punya keinginan sendiri dengan metode apa harus diperlakukan dan dianalisis. Lho, kok begitu? Ya, … memang begitu! Pernah mendengar istilah "psikologi data"? Tidak mudah memahami psikologi data seperti yang diperlihatkan antara lain oleh *swamping effect* dan *masking effect*. Sekelompok data tidak ada bedanya dengan sekelompok mahluk hidup. Ada anatominya. Ada strukturnya. Nah, struktur data inilah yang akan berbicara kepada kita dengan metode apa data itu "minta" diperlakukan dan dianalisis. Harap dicatat, "analisis data dan analisis statitik selalu diawali dengan dialog antara peneliti dengan data." Apa bahasa yang digunakan dalam dialog itu? Tidak ada bahasa yang lain, suka atau tidak suka, selain bahasa matematika.

Dialog pertama adalah tentang variabel WAKTU. Inilah yang dimaksud dengan langkah kanan dalam analisis data dan analisis statistik. Apakah pengambilan sampel terbebas dari pengaruh waktu

(*time independent*) atau tidak (*time dependent*)? Dengan kata lain, apakah pengaruh waktu turut diperhitungkan dalam pengambilan sampel? Nah, buku ini difokuskan pada kasus di mana pengaruh waktu tidak diperhitungkan dalam tujuan pengambilan sampel. Dalam kasus inilah digunakan "*The four plots*" atau Diagram-4 sebagai senjata utama pertama.

1. The Magnificent Seven

Bagaimana kalau pengaruh waktu turut diperhitungkan dalam pengambilan sampel? Topik ini adalah topik yang sangat luas yang mencakup analisis deret waktu (*time series analysis*) beserta aplikasinya. Pada kasus ini, data dikumpulkan dengan memperhitungkan pengaruh waktu. Jadi, ada urutan waktu dalam pengumpulan data. Setelah data terkumpul dan disajikan dalam urutan waktu (biasa disebut data deret waktu atau *time series data*), maka ada dua kemungkinan yang terjadi. Pertama, variabel waktu tidak berpengaruh terhadap data. Kedua, variabel waktu berpengaruh terhadap data. Kasus pertama adalah kasus yang kita jumpai umpamanya dalam pengendalian proses statistikal (*statistical process control*) disingkat PPS. Sedangkan kasus kedua adalah kasus yang kita hadapi dalam setiap analisis deret waktu (*time series analysis*) termasuk pemodelan deret waktu disingkat PDW dan peramalan (*forecasting*).

Dalam upaya memahami berbagai arsenal/persenjataan statistikal, urutan langkah berikut sangat kami rekomendasikan.

1. Mulailah dengan mengenal arsenal untuk kasus di mana pengaruh waktu tidak diperhitungkan dalam pengambilan sampel seperti yang kita bahas dalam buku ini.
2. Lanjutkan dengan mengenal arsenal untuk kasus di mana pengaruh waktu diperhitungkan dalam pengambilan sampel namun data terbebas dari pengaruh waktu.
3. Langkah terakhir adalah mengenal arsenal untuk kasus di mana pengaruh waktu diperhitungkan dalam pengambilan sampel dan data dipengaruhi oleh waktu.

Buku yang ada di tangan pembaca ini membahas "*The four plots*" atau Diagram-4 sebagai senjata utama pertama dalam arsenal yang disebut pada Butir 1 di atas. Adapun senjata utama pertama dalam

arsenal pada Butir 2 adalah "*The magnificent seven*" atau 7-teknik-jitu. Sedangkan untuk arsenal pada Butir 3 senjata utama pertamanya adalah prinsip Box-Jenkins yang akan dijelaskan pada bagian kedua di akhir bab ini.

Lalu, apa yang dimaksud dengan 7-teknik-jitu itu? Kaoru Ishikawa mengajarkan bahwa masalah peningkatan kualitas proses dapat diselesaikan dengan bantuan 7-teknik-jitu. Siapa Kaoru Ishikawa? Dia seorang guru di bidang sains kualitas di Jepang yang terkenal dengan diagram ciptaannya (diagram Ishikawa) dan yang mengembangkan sistem kualitas dan manajemen kualitas. Ketujuh teknik jitu itu adalah,

1. Histogram atau Stem and Leaf plot
2. Check sheet
3. Pareto chart
4. Cause-and-effect diagram
5. Defect concentration diagram
6. Scatter diagram
7. Control chart

Berikut deskripsi singkat dari ketujuh teknik itu.

1. Histogram atau *Stem and Leaf plot*; histogram sudah dibahas pada Sekapur Sirih, sedangkan *stem and leaf plot* identik dengan histogram dimana "*bin*" dalam histogram diganti dengan "*stem*" yakni satu atau beberapa digit(s) yang signifikan.
2. Check sheet; untuk mengidentifikasi berbagai informasi yang relevan tentang kualitas proses seperti misalnya kategori kegagalan proses, jenis bahan mentah proses, jenis produk, dll.
3. Pareto chart; untuk mengidentifikasi *the vital few* dan *the trivial many* dengan menggunakan aturan Pareto yang mengatakan: "masalah yang timbul dalam 80% populasi adalah akibat perilaku 20% populasi sisanya." Selanjutnya, *the vital few* digunakan untuk proses peningkatan kualitas.
4. Cause-and-effect diagram atau disebut juga diagram Ishikawa atau diagram *Fishbone*; berguna untuk mengidentifikasi faktor-faktor yang mempengaruhi kualitas proses.
5. Defect concentration diagram; berguna dalam mempelajari penyebab terjadinya kegagalan (*defects*).

6. Scatter diagram (diagram pencar); untuk menyelidiki hubungan antara 2 variabel kontinu (termasuk dalam pembuatan diagram Lag-1 yang dibahas di Sekapur Sirih).
7. Control chart; teknik yang paling kompleks di antara semua teknik dalam "*The magnificent seven*" mampu memberikan informasi yang diperlukan untuk mereduksi variabilitas proses dan sekaligus meningkatkan kualitas proses sepanjang waktu.

2. Prinsip Box-Jenkins

Prinsip Box-Jenkins adalah senjata utama pertama dalam melakukan analisis data deret waktu yang meliputi pemodelan dan peramalan. Prinsip ini terdiri atas tiga langkah berikut.

1. Identifikasi model; dengan melihat pola perilaku data deret waktu, dirumuskan model sementara yang dianggap cocok (dapat menggambarkan data).
2. Penaksiran parameter; semua parameter yang terlibat dalam model sementara tersebut ditaksir.
3. Validasi model; hasil penaksiran parameter kemudian digunakan untuk menguji kesahihan model. Apabila model tersebut dianggap sahih secara signifikan, maka model itu digunakan untuk analisis lebih lanjut. Apabila tidak signifikan, maka kembali ke langkah pertama dengan memodifikasi model awal. Demikian seterusnya sampai diperoleh model yang signifikan.

3. Prinsip Richard Feynman

"*You cannot make bricks without straw*" – begitu kata Winston Churchill (1904) sebagai Perdana Menteri Inggris Raya tatkala bicara tentang kebijakan publik. Dalam bahasa Bob-Jenkins, kata-kata Churchill itu menjadi: "Saya tidak akan dapat memahami alam tanpa ada data" seperti yang telah kita bahas di depan. Lain lagi dengan bahasa Richard Feynman. Kata Feynman: "Saya tidak akan dapat memahami alam tanpa kemampuan dan kejujuran intelektual."

Apabila prinsip Box-Jenkins dalam membangun model peramalan didasarkan kepada hasil analisis data deret waktu, Feynman mendasarkan pembangunan model itu pada kemampuan dan kejujuran intelektual tanpa harus ada data; data bukan yang utama. Dengan kata lain, apabila pendekatan Box-Jenkins bersifat induktif,

pendekatan Feynman bersifat deduktif. Dia berujar: "*What I cannot create, I do not understand*."

Dalam praktik, apa yang dilakukan Feynman terdiri atas 3 tahap berikut:

1. *Guess the law of nature*: Menentukan dugaan saintifik atau hipotesis mengenai hukum alam berdasarkan pengetahuan tentang cara kerja alam
2. *Deduce all the consequences*: Menyimpulkan apa yang harus terjadi jika dugaan itu benar. Setelah itu lalu melakukan verifikasi dengan membandingkan dugaan itu terhadap hasil eksperimen
3. *If it doesn't comply with nature, the guessed law is wrong*: Jika hasilnya tidak sesuai dengan eksperimen, maka dugaan mengenai hukum alam itu tidak benar. Gantilah dengan dugaan saintifik yang lain, lalu ulangi proses di atas.

Prinsip Feynman ini dapat digunakan untuk membuat model peramalan berdasarkan data deret waktu yang bernilai positif seperti biasa dihadapi oleh para ekonom khususnya para ahli bidang pasar keuangan (*financial market*). Untuk mengetahu bagaimana prinsip Feynman ini bekerja, silahkan baca buku karya Maman A. Djauhari dan Lee Siaw Li berjudul "*Forecasting Methods: The needs of policymakers*" terbitan UPM Press tahun 2018.

4. Penutup

Kemajuan dan perkembangan sains selalu dimulai dengan dugaan saintifik (hipotesis). Kemudian diikuti dengan berbagai perhitungan yang mengalir secara logis. Setelah itu, tahap yang sangat mendebarkan adalah menguji dugaan itu terhadap perilaku alam melalui eksperimen/observasi. Dikatakan mendebarkan karena eksperimen menjadi penentu utama kemajuan dan perkembangan sains. Dalam sains, *Experiment is King*. Sebuah teori harus sesuai dengan kenyataan; jika tidak, betapapun elegannya, teori itu salah. Ini adalah pesan kepada kita untuk tidak pernah menipu diri sendiri atau mempercayai otoritas secara membabi buta. Semuanya harus dipertanyakan termasuk ide-ide yang sudah mapan.

Pada dasarnya, siapa pun yang bergelut dengan sains harus memiliki jiwa yang "mengizinkan diri sendiri untuk ragu" artinya senantiasa

skeptis dan tidak puas (*disappointed*) dengan teori yang ada. Namun, kemudian harus diikuti dengan upaya untuk pengembangannya atau penyempurnaannya. Selain itu, harus juga dimiliki "kejujuran intelektual" umpamanya dengan tidak menyembunyikan data yang mungkin membuktikan bahwa teori yang kita buat salah. Atau dengan melaporkan semua fakta yang bertentangan dengan teori yang sedang kita bangun. Untuk kita catat bersama, yang namanya pengetahuan saintifik (*scientific knowledge*) adalah kumpulan pernyataan dengan tingkat kepastian yang bervariasi; tidak ada satupun yang benar-benar terbukti. Semua teori hanya didukung oleh bukti sementara. Pengetahuan saintifik bukan pengetahuan umum.

Akhir kata, buku ini kami tutup dengan mengutip pesan Albert Einstein sebagai berikut: "Jika kita tidak dapat menjelaskan sesuatu dengan sederhana, maka sebenarnya kita tidak cukup memahaminya." Kesederhanaan adalah keindahan ...!

BAB 15. EPILOG: KUMPULAN DATA

"*Data! data! data!" he cried impatiently. "I can't make bricks without clay*."

Arthur Conan Doyle

Untuk keperluan latihan, pada bab ini disajikan 6 (enam) buah kelompok data yang berasal dari berbagai sektor industri. Pada Tabel 15.1, nomor 1 adalah data dari industri makanan, lalu nomor 2 dan 3 dari industri keuangan, kemudian nomor 4-6 berturut-turut dari industri kesehatan, alat rumah tangga, dan curah hujan. Sedangkan kolom terakhir pada tabel itu menunjukkan nomor Apendiks pada buku "*Forecasting Methods: The needs of policymakers*" karya Maman A. Djauhari dan Lee Siaw Li (2018). Data orisinal dapat pembaca peroleh dari buku tersebut.

Tabel 15.1. Jenis data yang disiapkan untuk latihan

No.	Jenis Data	Apendiks
1	Kandungan lemak dalam adonan coklat (Jun 2011-Jul 2011)	5.1
2	Harga saham Petronas Dagangan Bhd harian (Jul 2015-Jul 2016)	3.6
3	Nilai tukar Ringgit Malaysia terhadap Dollar AS (Jul 2015-Jul 2016)	3.7
4	Banyaknya serangan penyakit cacar air mingguan (2012-2014)	3.5
5	Data angular dudukan sikat pada pembersih vakum (Apr 2012-Jun 2012)	3.3
6	Curah hujan tahunan (1878-1992)	2.2

Semua data orisinal merupakan data deret waktu di mana faktor waktu sangat berpengaruh (*time dependent*). Oleh karena itu, untuk membersihkan data dari kehadiran *outlier*, pengaruh waktu harus dihilangkan terlebih dahulu. Pada halaman-halaman selanjutnya, penulis sajikan data yang sudah terbebas dari pengaruh waktu dan siap dibersihkan dari kehadiran *outlier*.

Tabel 15.2. Data kandungan lemak dalam adonan coklat

No.	Data	No.	Data	No.	Data	No.	Data
1	-0,2641	26	0,0617	51	0,1395	76	-0,0644
2	0,0479	27	-0,0113	52	0,4370	77	0,0573
3	-0,0156	28	0,0536	53	-0,2520	78	0,0152
4	0,4513	29	-0,0463	54	-0,0413	79	0,0205
5	0,3339	30	-0,1259	55	-0,0928	80	0,1001
6	-0,2584	31	0,0680	56	-0,3428	81	-0,2922
7	0,1739	32	0,2222	57	0,3140	82	-0,0644
8	0,0328	33	0,1084	58	-0,1617	83	0,2556
9	-0,5469	34	0,2344	59	-0,1885	84	-0,1165
10	-0,0685	35	-0,1068	60	-0,1117	85	0,1951
11	-0,0724	36	-0,4810	61	-0,0404	86	-0,1195
12	-0,0933	37	-0,2173	62	0,3713	87	-0,1138
13	0,1204	38	0,0336	63	0,1670	88	0,2153
14	0,1579	39	0,1801	64	-0,1912	89	0,0228
15	0,1795	40	-0,1505	65	0,1094	90	0,0562
16	-0,1881	41	-0,1509	66	0,0736	91	0,1427
17	-0,2905	42	-0,0172	67	-0,0006	92	-0,0673
18	-0,0918	43	0,1402	68	0,1364	93	-0,2068
19	0,2996	44	0,0359	69	0,0816	94	-0,0776
20	-0,0721	45	-0,1864	70	0,0079	95	0,4112
21	-0,0410	46	-0,0437	71	-0,2461	96	-0,2209
22	0,0083	47	0,3605	72	0,2013	97	0,0250
23	0,1085	48	-0,1018	73	-0,0834	98	0,0318
24	0,0925	49	-0,0874	74	-0,2503		
25	-0,0932	50	0,1441	75	-0,0615		

Tabel 15.3. Harga harian saham Petronas Dagangan Bhd

No.	Data	No.	Data	No.	Data	No.	Data
1	0,0677	35	-0,0606	69	0,0458	103	-0,1150
2	-0,0443	36	-0,0187	70	-0,0482	104	0,0750
3	-0,0363	37	-0,1727	71	-0,1382	105	0,0351
4	0,1057	38	-0,1086	72	-0,0662	106	-0,2710
5	-0,0003	39	0,0594	73	-0,1781	107	-0,0809
6	-0,0324	40	0,4554	74	-0,0700	108	0,0211
7	0,0256	41	-0,0262	75	-0,0180	109	-0,0909
8	-0,0284	42	0,0011	76	-0,0340	110	-0,1428
9	-0,0944	43	-0,0108	77	-0,1960	111	-0,4068
10	-0,0004	44	0,0471	78	-0,3719	112	-0,0524
11	0,0097	45	-0,0269	79	-0,3076	113	-0,0145
12	-0,0104	46	0,2251	80	-0,0395	114	-0,1345
13	0,0276	47	0,0312	81	-0,0135	115	-0,1065
14	0,0116	48	-0,0110	82	0,0465	116	-0,1224
15	0,4096	49	-0,0130	83	-0,0075	117	-0,0244
16	0,1499	50	0,0070	84	-0,0135	118	0,0857
17	0,0394	51	-0,0510	85	0,5865	119	-0,0043
18	0,1113	52	0,0030	86	0,0271	120	-0,0144
19	0,0793	53	0,0290	87	-0,0159	121	0,9856
20	-0,1648	54	-0,0090	88	0,9261	122	0,1273
21	-0,0288	55	0,1869	89	-0,5783	123	-0,1910
22	-0,7128	56	-0,2330	90	-0,0795	124	-0,0329
23	-0,3416	57	0,0430	91	0,0059	125	-0,0149
24	-0,0582	58	0,0150	92	-0,1121	126	-1,2349
25	-0,2942	59	0,1689	93	-0,0240	127	0,7260
26	0,0000	60	0,1449	94	0,6460	128	0,0731
27	0,0720	61	0,0408	95	0,2527	129	-0,7947
28	0,0360	62	0,1707	96	0,8056	130	0,1885

29	-0,3681	63	0,7647	97	0,1065	131	-0,0861
30	0,4522	64	0,0636	98	-0,0110	132	0,0156
31	0,0786	65	0,3661	99	-0,1150	133	0,0497
32	0,5118	66	-0,0156	100	0,0350	134	-0,0084
33	0,2603	67	0.1619	101	-0,0090	135	-0,1144
34	0,1295	68	0,0040	102	-0,0150	136	-0,3644

Tabel 15.3 (Sambungan)

No.	Data	No.	Data	No.	Data	No.	Data
137	0,1119	171	0,2316	205	-0,0283	239	-0,2063
138	0,0819	172	0,0355	206	0,2637	240	-0,0942
139	-0,0062	173	-0,1329	207	-0,2262	241	-0,3802
140	0,4257	174	-0,0868	208	-0,0783	242	0,1101
141	0,0300	175	-0,1008	209	0,0217	243	-0,0179
142	0,1255	176	-0,4227	210	-0,0103	244	0,0239
143	0,0395	177	0,0456	211	-0,0143	245	-0,1300
144	0,3094	178	-0,0445	212	-0,0143	246	0,0340
145	0,8175	179	0,0615	213	-0,1143	247	0,1120
146	0,0662	180	0,0535	214	-0,2043	248	-0,0020
147	0,3247	181	-0,0086	215	-0,1922	249	-0,1141
148	-0,6212	182	-0,1146	216	-0,4301	250	0,0159
149	0,2813	183	-0,1046	217	-0,1937	251	-0,2300
150	-0,4790	184	0,0375	218	-0,1877	252	-0,0359
151	-0,0649	185	0,0115	219	-0,0096	253	0,1261
152	-0,0151	186	-0,0125	220	-0,0316	254	0,0401
153	-0,2151	187	-0,2145	221	0,0044	255	0,0700
154	0,0650	188	0,0656	222	0,5484	256	-0,0260
155	-0,0049	189	-0,0044	223	0,0829	257	0,0639
156	0,0650	190	-0,0145	224	-0,0100	258	-0,0061
157	-0,0470	191	-0,1745	225	0,0660	259	-0,0141
158	0,3410	192	-0,0704	226	0,2340	260	-0,1141
159	0,0212	193	-0,0584	227	-0,1100		
160	-0,0152	194	-0,0183	228	-0,0261		
161	-0,0752	195	-0,0143	229	0,1059		
162	-0,1612	196	-0,0343	230	0,0579		
163	-0,1292	197	-0,0163	231	-0,0682		
164	-0,1851	198	0,0057	232	-0,2202		

165	0,1690	199	-0,0123	233	0,0659
166	-0,4349	200	0,0857	234	0,0359
167	0,2213	201	-0,1043	235	0,0099
168	0,0333	202	-0,1243	236	0,1478
169	-0,0330	203	-0,0243	237	0,0419
170	-0,5570	204	0,0457	238	0,0697

Tabel 15.4. Nilai tukar Ringgit Malaysia terhadap Dollar AS

No.	Nilai	No.	Nilai	No.	Nilai	No.	Nilai
1	-0,0004	32	-0,0071	63	-0,0153	94	0,0547
2	0,0180	33	0,0469	64	-0,0014	95	-0,0092
3	-0,0069	34	0,0395	65	0,0008	96	-0,0622
4	-0,0010	35	-0,0014	66	-0,0353	97	0,0449
5	-0,0097	36	-0,0009	67	-0,0576	98	0,0063
6	0,0009	37	0,0242	68	-0,0051	99	-0,0006
7	0,0067	38	-0,0089	69	0,0076	100	0,0834
8	-0,0005	39	0,0148	70	0,0052	101	0,0269
9	0,0075	40	0,0086	71	-0,0006	102	0,0146
10	-0,0369	41	0,0677	72	0,0377	103	-0,0428
11	0,0276	42	0,0043	73	0,0471	104	-0,0114
12	0,0115	43	-0,0007	74	0,0684	105	-0,0012
13	0,0036	44	0,0522	75	0,0332	106	-0,0009
14	-0,0018	45	-0,0213	76	-0,0122	107	-0,0156
15	-0,0008	46	0,0262	77	0,0009	108	0,0443
16	-0,0036	47	-0,0479	78	-0,0008	109	-0,0266
17	-0,0154	48	-0,0199	79	0,0276	110	0,0806
18	-0,0018	49	-0,0030	80	0,0351	111	-0,0327
19	0,0247	50	-0,0009	81	-0,0738	112	-0,0427
20	0,0007	51	0,0227	82	0,0373	113	-0,0024
21	0,0003	52	-0,0038	83	-0,0120	114	0,0434
22	-0,0007	53	0,0259	84	0,0058	115	-0,0485
23	0,0247	54	0,0166	85	-0,0006	116	0,0341
24	0,0097	55	0,0628	86	-0,0754	117	0,0172
25	0,0236	56	0,0140	87	-0,0419	118	0,0598
26	0,0219	57	-0,0004	88	-0,1151	119	-0,0121
27	0,0210	58	0,0301	89	-0,0100	120	-0,0014
28	-0,0103	59	-0,0299	90	-0,0510	121	0,0316

29	-0,0012	60	-0,0018	91	-0,0047	122	-0,0171
30	0,0005	61	-0,0149	92	-0,0009	123	-0,0225
31	0,0695	62	0,0191	93	0,0007	124	0,0218

Tabel 15.4 (Sambungan 1)

No.	Nilai	No.	Nilai	No.	Nilai	No.	Nilai
125	0,0193	156	0,0206	187	0,0228	218	-0,0014
126	0,0011	157	-0,0376	188	0,0319	219	-0,0183
127	-0,0008	158	-0,0123	189	-0,0005	220	0,0650
128	-0,0069	159	0,0093	190	-0,0009	221	-0,0303
129	0,0001	160	-0,0405	191	-0,0306	222	0,0130
130	-0,0345	161	0,0046	192	-0,0250	223	0,0332
131	-0,0743	162	-0,0006	193	0,0315	224	-0,0068
132	-0,0149	163	0,0016	194	-0,0626	225	-0,0011
133	0,0002	164	0,0197	195	-0,0762	226	-0,0251
134	-0,0008	165	-0,0201	196	-0,0016	227	0,0362
135	0,0081	166	0,0004	197	-0,0008	228	0,0146
136	-0,0655	167	-0,0013	198	-0,0031	229	-0,0153
137	-0,0135	168	-0,0009	199	-0,0114	230	-0,0331
138	0,0198	169	-0,0009	200	-0,0312	231	0,0047
139	0,0406	170	0,0035	201	-0,0645	232	-0,0006
140	-0,0008	171	-0,0141	202	-0,0210	233	-0,0037
141	-0,0009	172	0,0208	203	-0,0070	234	-0,0395
142	-0,0191	173	0,0043	204	-0,0010	235	-0,0122
143	-0,0293	174	-0,0082	205	0,0167	236	-0,0148
144	0,0244	175	-0,0011	206	0,0591	237	-0,0743
145	-0,0169	176	-0,0009	207	-0,0286	238	0,0005
146	-0,0408	177	0,0541	208	-0,0943	239	-0,0007
147	-0,0068	178	-0,0043	209	0,0552	240	0,0125
148	-0,0010	179	0,0699	210	0,0052	241	0,0504
149	0,0695	180	-0,0492	211	-0,0007	242	-0,0299
150	0,0143	181	0,0412	212	0,0204	243	0,0255
151	-0,0158	182	0,0073	213	-0,0272	244	-0,0548
152	0,0022	183	-0,0006	214	-0,0322	245	0,0019

153	0,0608	184	-0,0239	215	0,0100	246	-0,0006
154	0,0165	185	0,0013	216	0,0286	247	0,0517
155	-0,0003	186	-0,0348	217	-0,0152	248	0,0312

Tabel 15.4 (Sambungan 2)

No.	Nilai	No.	Nilai	No.	Nilai	No.	Nilai
249	0,0183	280	0,0002	311	-0,0028	342	-0,0135
250	-0,1412	281	-0,0008	312	0,0301	343	-0,0096
251	0,0142	282	-0,0058	313	0,0359	344	-0,0011
252	-0,0010	283	-0,0543	314	-0,0090	345	-0,0313
253	-0,0008	284	0,0175	315	0,0006	346	-0,0215
254	-0,0143	285	0,0503	316	-0,0007	347	-0,0045
255	-0,0540	286	0,0001	317	0,0112	348	-0,0601
256	0,0299	287	-0,0053	318	0,0158	349	0,1145
257	0,0059	288	-0,0010	319	-0,0057	350	0,0062
258	0,0004	289	0,0047	320	-0,0330	351	-0,0007
259	0,0047	290	0,0034	321	0,0130	352	0,0287
260	-0,0006	291	-0,0029	322	0,0067	353	-0,0662
261	-0,0168	292	-0,0364	323	-0,0006	354	-0,0466
262	-0,0424	293	0,0378	324	0,0220	355	-0,0196
263	-0,0513	294	-0,0007	325	-0,0008	356	-0,0027
264	-0,0551	295	-0,0008	326	0,0384	357	-0,0028
265	0,0098	296	0,0076	327	-0,0006	358	-0,0009
266	0,0077	297	0,0541	328	-0,0591	359	-0,0008
267	-0,0005	298	0,0493	329	-0,0123	360	0,0422
268	0,0122	299	-0,0130	330	-0,0012	361	0,0111
269	0,0387	300	-0,0198	331	-0,0235	362	-0,0074
270	-0,0565	301	0,0000	332	-0,0082	363	-0,0409
271	0,0533	302	-0,0007	333	-0,0279	364	0,0022
272	-0,0299	303	0,0547	334	0,0187		
273	0,0038	304	-0,0087	335	0,0395		
274	-0,0006	305	-0,0347	336	-0,0053		
275	-0,0394	306	0,0137	337	-0,0010		
276	-0,0167	307	0,0075	338	0,0017		

277	0,0152	308	0,0053	339	0,0218
278	0,0065	309	-0,0006	340	-0,0128
279	0,0123	310	-0,0193	341	0,0115

Tabel 15.5. Data mingguan banyaknya serangan penyakit cacar air

No.	Data	No.	Data	No.	Data	No.	Data
1	-2,4108	37	2,0376	73	-2,8660	109	3,2864
2	6,5280	38	-0,8661	74	-0,9010	110	-1,4108
3	-1,3108	39	0,0183	75	3,5695	111	2,0219
4	-0,8980	40	-0,4228	76	-1,2826	112	-1,2992
5	0,1390	41	-0,4806	77	2,5444	113	-0,4726
6	1,5339	42	2,0376	78	3,3167	114	-5,4133
7	-2,8594	43	-2,8661	79	-0,4205	115	0,1723
8	0,0445	44	1,0762	80	0,5216	116	5,7087
9	-2,8624	45	-1,2875	81	5,5795	117	-1,7181
10	4,0512	46	-3,4606	82	-2,1269	118	-0,9488
11	-0,5250	47	0,0210	83	0,5850	119	3,5818
12	1,0445	48	0,5695	84	5,7023	120	-0,2987
13	-0,2992	49	2,0290	85	-3,7270	121	-1,9781
14	0,0234	50	3,1349	86	-3,8809	122	5,5264
15	2,5892	51	6,1337	87	-0,4760	123	-1,5354
16	6,1406	52	-4,1189	88	10,0434	124	-2,4278
17	-2,1551	53	1,7421	89	0,9751	125	-4,9508
18	2,5563	54	-0,0984	90	-1,4407	126	-1,3861
19	-2,1551	55	1,5306	91	-3,4057	127	1,5750
20	-4,4437	56	2,7038	92	2,0445	128	-2,4228
21	-4,9750	57	3,5867	93	-0,7246	129	3,5394
22	-2,9488	58	-4,2995	94	-3,4728	130	0,9149
23	5,5216	59	-1,8704	95	2,0207	131	-1,9812
24	3,4869	60	-1,4182	96	-4,2987	132	-0,4806
25	1,0579	61	2,0188	97	0,1559	133	3,0376
26	-4,4057	62	-2,2962	98	-0,2941	134	0,7059
27	-3,4219	63	3,0612	99	-3,9812	135	-0,4760
28	2,5240	64	-0,1095	100	-2,4606	136	-2,9566
29	-0,2977	65	-2,4710	101	-2,4867	137	0,5356
30	-0,9793	66	1,5216	102	7,0237	138	1,1337
31	-2,9537	67	-1,8656	103	7,9436	139	-1,9595
32	4,5327	68	6,5327	104	-4,2987	140	0,0181
33	-4,4992	69	-2,2473	105	-0,8441	141	2,5750

34	-0,6453	70	0,7010	106	1,1339
35	0,7125	71	-4,7136	107	-0,9566
36	-0,4806	72	-1,2323	108	3,5216

Tabel 15.6. Data angular lapisan dudukan sikat pada pembersih vakum

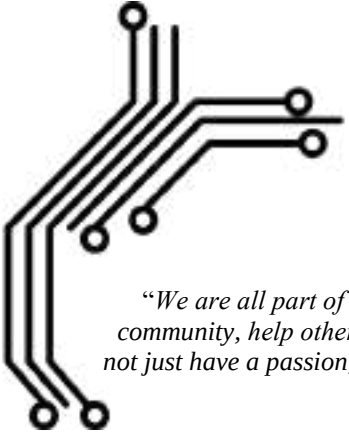
No.	Data	No.	Data	No.	Data	No.	Data
1	-0,0246	26	-0,0191	51	-0,0159	76	-0,0069
2	0,0137	27	0,0468	52	0,0071	77	0,0207
3	-0,0068	28	-0,0065	53	-0,0243	78	0,0246
4	0,0407	29	0,0165	54	-0,0024	79	0,0007
5	-0,0180	30	-0,0187	55	0,0111	80	-0,0279
6	0,0369	31	0,0149	56	-0,0221	81	-0,0446
7	-0,0052	32	-0,0038	57	-0,0343	82	-0,0125
8	-0,0133	33	0,0229	58	-0,0079	83	0,0075
9	-0,0176	34	-0,0023	59	-0,0332	84	0,0147
10	-0,0589	35	-0,0171	60	0,0151	85	-0,0679
11	-0,0093	36	-0,0459	61	0,0223	86	0,0262
12	0,0258	37	-0,0466	62	-0,0044	87	0,0512
13	-0,0411	38	-0,0105	63	0,0268	88	0,0146
14	-0,0185	39	0,0656	64	0,0220		
15	0,0111	40	0,0079	65	0,0155		
16	0,0359	41	0,0294	66	0,0201		
17	-0,0265	42	-0,0329	67	0,0164		
18	0,0133	43	0,0040	68	-0,0138		
19	-0,0269	44	0,0096	69	-0,0048		
20	0,0104	45	-0,0232	70	-0,0361		
21	0,0069	46	0,0077	71	0,0039		
22	0,0585	47	-0,0072	72	0,0066		
23	0,0003	48	0,0267	73	-0,0284		
24	-0,0121	49	0,0100	74	0,0064		
25	0,0254	50	0,0147	75	0,0147		

Tabel 15.7. Data curah hujan tahunan

No.	Data	No.	Data	No.	Data	No.	Data
1	-0,4907	31	-13,0554	61	1,8537	91	2,7482
2	-12,4792	32	6,6805	62	15,7961	92	-11,8143
3	0,3287	33	0,6665	63	-17,5785	93	-5,9888
4	6,5279	34	3,7883	64	6,9028	94	9,6079
5	28,0238	35	10,3849	65	4,7986	95	6,2044
6	-12,8674	36	-3,1900	66	-7,1829	96	-6,3945
7	-4,4489	37	3,4748	67	1,1820	97	-2,4824
8	2,8616	38	-11,1395	68	-10,2113	98	4,1136
9	4,4268	39	3,1655	69	-0,8261	99	17,7994
10	15,0770	40	-3,1727	70	5,0938	100	-4,1089
11	-13,4028	41	-1,4198	71	-1,5497	101	3,2484
12	-8,1168	42	9,9585	72	5,4015	102	-10,0711
13	5,9480	43	0,5295	73	14,7924	103	-2,8584
14	6,5593	44	-11,2647	74	-14,6371	104	21,5545
15	-12,7176	45	-1,8415	75	3,2239	105	-12,8913
16	-0,5784	46	1,8514	76	4,5875	106	-8,9684
17	2,1822	47	10,0547	77	0,8158	107	9,0833
18	2,0862	48	5,9236	78	0,5384	108	-3,3963
19	-8,1123	49	-9,9185	79	4,0431	109	-1,4197
20	0,1981	50	-4,6077	80	-8,9158	110	-5,6663
21	4,8911	51	4,5029	81	-1,0926	111	-0,9146
22	2,0941	52	8,6056	82	-1,8322	112	9,6635
23	1,4984	53	-4,8759	83	7,8208	113	12,9162
24	2,2564	54	4,3418	84	3,0164		
25	-2,0191	55	0,6338	85	-5,8468		
26	5,8757	56	-2,1937	86	16,8053		
27	6,4918	57	3,6450	87	-1,3815		
28	-4,9646	58	1,7737	88	4,7695		
29	-4,5165	59	9,0357	89	-9,7026		

30	9,3784	60	-9,8698	90	12,6119

Sebagai catatan, khususnya bagi mereka yang tertarik dalam menghilangkan pengaruh waktu pada data orisinal, teknik *geometric Brownian motion* (GBM) adalah teknik yang sangat penulis rekomendasikan tatkala semua data deret waktu bernilai positif.



REFERENSI

"*We are all part of the flow of history … that is going to help the community, help other people … so people will say, this person did not just have a passion, he cared about making something that other people could benefit from.*"

Steve Jobs

1. Adikaram, K.K.L.B., Hussein, M.A., Effenberger, M., dan Becker, T. (2015). Data transformation technique to improve the outlier detection power of Grubbs' test for data expected to follow linear relation. *Journal of Applied Mathematics*. http://dx.doi.org/10.1155/2015/708948
2. Anderson, T.W., dan Darling, D.A. (1952) Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematica Statistics*, 23: 193-212.
3. Anscombe, F.J. (1973). Graphs in Statistical Analysis. *American Statistician*, **27**(1): 17-21. doi:10.1080/00031305.1973.10478966.
4. Anscombe, F.J., and Guttman, I. (1960). Rejection of outliers. *Technometrics*, 2(2): 123-147.
5. Barbato, G., Barini, E.M., Genta, G., & Levi, R. (2011). Features and performance of some outlier detection methods. *Journal of Applied Statistics*, 38(10): 2133-2149.
6. Barnett, V., and Lewis, T. (1994), *Outliers in statistical data*. Third edition. Chichester, John Wiley, NY.
7. Benzécri, J-P. (1960). *Sur les variétés localement affines et localement projectives*. Bulletin de la S. M. F., 88, p. 229-332.
8. Bishop, C.M. (August 1994). Novelty detection and neural network validation. IEE Proceedings - Vision, Image, and Signal Processing, 141(4): 217–222.
9. Bohrer, A. (2008). One-sided and two-sided critical values for Dixon's outlier test for sample sizes up to n = 30. *Economic Quality Control*, 23(1): 5-13.

10. Bolton, S. (1990). *Pharmaceutical statistics: Practical and clinical application*. Drug and the pharmaceutical sciences, 2nd Edition. Marcel Dekker, Inc., New York.
11. Caillez, F., dan Pages, J-P. 1976). *Analyse des Données*. SMASH (Societé de Mathématiques Appliquées et de Sciences Humaines), Paris.
12. Chatterjee, S., dan Hadi, A.S. (2006). *Regression Analysis by Example*. John Wiley and Sons. ISBN 0-471-74696-7.
13. Chrominski, K., dan Magdalena, TKACZ. (2010). Comparison of outlier detection methods in biomedical data. *Journal of Medical Informatics & Technologies*, 16, ISSN 1642-6037.
14. Cleveland, W.S. (2001). Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistical Review*, 69(1): 21-26. https://www.jstor.org/stable/1403527
15. Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19(1): 15–18.
16. de Tocqueville, A. (1909). *De la démocratie en Amérique*, Tome 2, Deuxième partie, Chapitre 5,
17. Dixon, W.J., (1950), Analysis of extreme values. *Annals of Mathematical Statistics*, 21(4): 488–506.
18. Dixon, W.J. (1960). Simplified estimation from censored normal samples. *The Annals of Mathematical Statistics*, 31(2): 385–391.
19. Djauhari, M.A. (2001a). Improving the ESD procedure for outlier testing. *BioPharm: The applied technologies of biopharmaceutical development*, March 2001, Milwaukee, USA.
20. Djauhari, M.A. (2001b). Letter to the Editor: Deriving the distribution. *Biopharm: The applied technologies of biopharmaceutical development*, July 2001, Milwaukee, USA.
21. Djauhari, M.A. (2003). Statistical testing for outliers: Calculating the critical point of the extreme studentized deviation using the beta inverse function. *BioPharm International*, October 2003, Milwaukee, USA.
22. Djauhari, M.A. (2007). A measure of multivariate data concentration. *Journal of Applied Probability and Statistics*, 2(2): 139–155.
23. Djauhari, M.A. (2020). *Ukuran sampel: Formula generik bagi praktisi sains sosial*. ITB Press, ISBN: 978-623-7568-90-2.

24. Djauhari, M.A. (2021). *Teknik membersihkan data: Awas ... outlier ... Outlier ... OUTLIER*! ITB Press, ISBN: 978-623-297-164-6.
25. Djauhari, M.A., dan Lee, S.L. (2018). *Forecasting Methods: The needs of policymakers*. UPM Press, 2018.
26. Erickson, B.H., dan Nosanchuk, T.A. (1979). *Understanding data*. Milton Keynes: Open University Press.
27. Escoufier Y., Fichet, B., Lebart, L., Hayashi, C., Ohsumi, N., dan Baba, Y. (Eds.) (1995), "*Data Science and Its Applications*", Academic Press, Tokyo, Jepang.
28. Everitt, B.S. (2002). *Cambridge Dictionary of Statistics*. Cambridge University Press. ISBN 0-521-81099-X.
29. García, S., Luengo, J., dan Herrera, F. (2015). *Data Preprocessing in Data Mining*. Springer Cham Heidelberg, New York. ISBN 978-3-319-10247-4 (eBook) DOI 10.1007/978-3-319-10247-4.
30. Gladwell, M. (2008). *Outliers: The story of success*. Little, Brown & Company, NY.
31. Greenacre, M. (2007). *Correspondence Analysis*. 2nd Edition, Chapman & Hall/CRC. ISBN -10: 1-58488-616-1.
32. Grubbs, F.E. (1950). Sample criteria for testing outlying observations. *Annales of Mathematical Statistics*, 21(1): 27-58.
33. Grubbs F.E. (1969), Procedures for detecting outlying observations in samples. American Statistical Association and American Society for Quality. *Technometrics*, 11(1): 1-21.
34. Grubbs, F.E., dan G. Beck (1972), Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics*, 14(4): 847–854.
35. Han, J., Kamber, M., dan Pei, J. (2012). *Data Mining Concepts and Techniques*, 3rd Edition. Morgan Kaufmann Publishers (an imprint of Elsevier), Waltham, MA 02451, USA.
36. Hardin, J., dan Rocke, D.M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14(4): 928-946.
37. Hawkins D.M. (1980), *Identification of outliers*. Chapman & Hall, London.
38. Hayashi, C. (1951). *On the Prediction of Phenomena from Qualitative Data and the Quantification of Qualitative Data from*

the Mathematico-Statistical Point of View. The Institute of Statistical Mathematics, Tokyo.
39. Hayashi, C. (1998). *What is Data Science? Fundamental Concepts and a Heuristic Example*. In: Hayashi, C., Yajima, K., Bock, HH., Ohsumi, N., Tanaka, Y., Baba, Y. (eds) Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Tokyo. https://doi.org/10.1007/978-4-431-65950-1_3.
40. Herwindiati, D.E., Djauhari, M.A., dan Mashuri, M. (2007). Robust multivariate outlier labelling. *Communication in Statistics-Simulation and Computation*, 36(6): 1287-1294.
41. Hodge, V.J., dan Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2): 85-126.
42. Iglewicz, B., dan Hoaglin, D.C. (1993). How to detect and handle with outliers. *American Society for Quality*, Statistics Division.
43. Irwin, J.O. (1925). On a criterion for the rejection of outlying observations. *Biometrika*, 17: 238-250.
44. Jain R.B., and Pingel, L.A. (1981). A procedure for estimating the number of outliers. *Communications in Statistics – Theory and Methods*, 10: 1029–1041.
45. Kumar, S. (2021). Handle outliers using logistic regression. Dapat diunduh dari sumber berikut, https://medium.com/geekculture/essential-guide-to-handle-outliers-for-your-logistic-regression-model-63c97690a84d
46. Liu, H., Shah, S., dan Jiang (2014), On-line outlier detection and data cleaning. *Computation in Chemical Engineering*, 28: 1635–1647.
47. Lohr, S.L. (1999). *Sampling: Design and analysis*. Duxbury Press, London.
48. Maddala, G.S. (1992). *Outliers: Introduction to econometrics*, 2nd Edition. MacMillan, NY.
49. Manoj, K., and Kannan, K.S. (2013). Comparison of methods for detecting outliers. *International Journal of Scientific & Engineering Research*, 4(9): 709-714.
50. Marsh, C. (1988). *Exploring data: An introduction to data analysis for social scientists*. Polity Press, Cambridge, UK.
51. Newman, M.E.J. (2006). Power laws, Pareto distributions and Zipf's law. https://arxiv.org/PS_cache/cond-mat/pdf/0412/0412004v3.pdf

52. NIST/SEMATECH: *Engineering statistics handbook*. Dapat diunduh dari sumber berikut, https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm dan https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm
53. Nosanchuk, T.A., dan Erickson, B.H. (1992). *Understanding data*, 2nd Edition. ASIN: B0032OM19O, Publisher: University of Toronto Press, Canada.
54. Ohsumi, N. (2004). *Memories of Chikio Hayashi and his great achievement*. Diunduh pada 11 September 2025 dari https://www.researchgate.net/publication/268719699
55. Pareto, V. (1909). *Manuel d'Economie Politique* (diterjemahkan dari bahasa Itali ke bahasa Prancis oleh Alfred Bonnet). V. Giard & E. Briere Libraires-Editeurs, Paris 5e. http://digamoo.free.fr/pareto1909.pdf
56. Paul, S. R., and Fung, K.Y. (1991). A generalized extreme studentized residual multiple-outlier detection procedure in linear regression. *Technometrics*, 33(3): 339-348.
57. Peirce, B. (1852). Criterion for the rejection of doubtful observations. *Astronomical Journal* II 45.
58. Pope, A. (1976). The statistics of residuals and the detection of outliers. NOAA Technical Report, NOS 65, NGS 1 Rockville, MD.
59. Quesenberry, C.P., and David, H.A. (1961). Some tests for outliers. *Biometrika*, 48(3/4): 379-390.
60. Razali, N., dan Wah, B. (2011). Power comparison of Shaphiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling tests. *Journal of Statistical Modelling and Analytics*, 2: 21-33.
61. Rosner, B. (1975), On the detection of many outliers. *Technometrics*, 17(2): 221-227.
62. Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25(2): 165-172.
63. Rousseeuw, P. (1985). *Multivariate estimation with high breakdown point*. In Mathematical Statistics and Applications, Vol. B, Eds. W. Grossman, G. Pflug, I, Vincze, and W. Wertz, Dordrecht: Reidel, p. 283-297.
64. Rousseeuw, P., dan Leroy, A. (1996). *Robust regression and outlier detection*, 3rd Edition, John Wiley & Sons, NY.

65. Rousseeuw, P., dan van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41: 212-223.
66. Ruiz, A.Q., dan Verma, S.P. (2006). Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering. *Revista Mexicana de Ciencias Geologica*, 23(2): 133-161.
67. Saville, D.J., dan Wood, G.R. (1991). *Statistical Methods: The geometric approach*. Springer. ISBN 0-387-97517-9.
68. Thompson, R. (1985). A note on restricted maximum likelihood estimation with an alternative outlier model. *Journal of the Royal Statistical Society, Series B*, 47(1): 53-55
69. Tietjen G.L., Moore R.H., Beckman R.J. (1973). Testing for a single outlier in simple linear regression. *Technometrics*, 15(4): 717- 721.
70. Tietjen, G.L., dan Moore, R.H. (1972). Some Grubbs-type statistics for the detection of several outliers. *Technometrics*, 14(3): 583–597.
71. Tufte, E.R. (2001). *The visual display of quantitative information*, 2nd Ed. Cheshire, CT. Graphics Press. ISBN 0-9613921-4-2.
72. Tukey, J. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company Reading, London.
73. Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wisley Publishing Company, Inc. ISBN 0-201-07616-0.
74. Weisberg, S. (1985). *Applied linear regression*, 2nd Edition. John Wiley & Sons, NY.
75. Woolley, T.W. Jr (2010). The effect of swamping on outlier detection in normal samples. Joint Statistical Meetings - Biometrics Section-to include ENAR & WNAR pp. 3793-3798
76. Zerbet, A., dan Nikulin, M. (2003). A new statistic for detecting outliers in exponential case. *Communications in Statistics - Theory and Methods*, 32: 573–583.
77. Zimek, A., Schubert, E., dan Kriegel, H.P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*. 5(5): 363–387.

INDEKS SUBJEK

Give us the tool and we will finish the job."

Winston Churchill

PENGHARGAAN DAN TERIMA KASIH

Penampilan buku yang atraktif dan menarik, baik penampilan ilustrasi jilid muka dan belakang maupun pengaturan tata ruang dan tata letak, diharapkan akan lebih menggariahkan para pembaca. Buku ini menawarkan pengalaman petualangan intelektual yang unik dan tak terlupakan dalam membersihkan data. Di dalamnya disajikan 10 teknik utama yang perlu diketahui oleh para profesional (calon profesional) tatkala membersihkan data dari kehadiran outlier.

Sapuan serta polesan yang menghasilkan penampilan yang atraktif itu terwujud berkat tangan-tangan terampil para profesional di Penerbit Campustaka, anggota IKAPI No. 635/DKI/2024, beralamat di Jl. H. Lebar No. 21B, RT 02/RW 01, Meruya Selatan, Kembangan, Jakarta Barat 11650. Mereka layak mendapat penghargaan atas upaya menyajikan buku ini kepada para pembaca baik Dosen, peneliti, birokrat, politisi, maupun khalayak ramai, yang hendak terjun menekuni seluk beluk analisis data dan analisis statistik.

Atas kerja keras kawan-kawan di Penerbit Campustaka, kami para penulis mengucapkan terima kasih dan penghargaan yang tinggi.

TENTANG PARA PENULIS

1. Maman Abdurachman Djauhari

Maman Abdurachman Djauhari lahir di Dayeuh Garut, Tanah Pasundan, pada 8 Desember 1948. Dia dianugrahi Medali Emas "*for Outstanding Contribution to Statistical Science*" oleh ISOSS (Islamic Countries Society of Statistical Sciences), Desember 2005.

Docteur du troisiéme cycle dalam bidang Mathématique Pure et Appliquées dan Diplôme d'Études Approfondies (DEA), diraihnya di Université de Montpellier 2 (Université Science et Technique du Languedoc), Prancis, pada tahun 1979 dan 1977. Sebelumnya, Doctorandus dan Sarjana Satu (sekarang S1) dalam bidang Matematika diraih di Institut Teknologi Bandung (ITB), tahun 1974 dan 1973.

Saat ini Maman berkhidmat sebagai Dosen di STAI KH. Badruzzaman, Garut. Berikut adalah jejak-jejak terpenting perjalanan karirnya.

Sebagai Penulis Internasional (data per Desember 2025):

1. Penulis 86 buah artikel di jurnal Scopus; 462 kali sitasi, dengan *h*-index 10.
2. Reviewer untuk tidak kurang dari 25 jurnal internasional
3. Penulis utama buku "*Reliable Shewhart-type control charts for multivariate process variability*" (bersama D.E. Herwindiati dan Z. Zolkeply) yang terbit di Jerman, 2018. Pada Juni 2020 buku tersebut telah diterbitkan ulang dalam 8 Bahasa Eropa (Belanda, Itali, Jerman, Perancis, Polandia, Portugis, Rusia, dan Spanyol).
4. Penulis utama buku "Forecasting Methods: The needs of policymakers" (Bersama Lee Siaw Li), UPM Press, 2018.
5. Penulis buku "Advanced monitoring techniques for complex process variability in manufacturing industry" UPM Press, 2016.

Karir Profesional Internasional:

1. Konsultan pada Bank Dunia (dalam bidang kualitas data), 2009.
2. Anggota Forum Eksekutif, Coleman Research Group, New York, 2006.
3. Konsultan dalam analisis statistik bagi para peneliti di 9 lembaga internasional (6 di Amerika Serikat, 1 di India, 1 di Hong Kong, dan 1 di Polandia), 2000.
4. Anggota pendiri International Statistics Forum, Taejon, Korea Selatan, 1999.

Pengakuan Internasional:

1. Professor pada Harvard Business School-UTM Twin Program, 2010-2013.
2. Scientific Board, wakil Indonesia, pada APEC (Asia-Pacific Econophysics Conference), sejak 2016.
3. Pengakuan dari UNESCO dan juga dari AEGIS (Australian Experts Group in Industrial Study) sebagai aktivis dalam memproduksi *scientific knowledge* di Asia-Pasifik, 2005.
4. Professor di UNIMAP (2008), UTM (2009-2014), dan UPM (2014-2016).
5. Academic Advisor di Universiti Teknologi MARA, Shah Alam, Malaysia, dan juga di Universiti Tun Hussein Onn Malaysia, tahun 2012-2014.

Karir Akademik:

1. Ketua Majelis Guru Besar, ITB, 2007-2008.
2. Sekretaris Majelis Guru Besar, ITB, 2001-2007.
3. Dekan, FMIPA, ITB, 1997-2000.

Penghargaan:

1. Satya Lencana Karya Satya XXX dari Presiden Republik Indonesia, 2007
2. Excellent Service Award, Faculty of Science, Universiti Teknologi Malaysia, 2013.
3. Award in Consultation, Universiti Teknologi Malaysia, 2011.
4. Dosen Terbaik, Departmen Matematika, Institut Teknologi Bandung, 2000.

Minat Dalam Riset (diurut abjad):

1. Complex Networks Analysis
2. Financial Market Analysis
3. Mathematical Statistics
4. Multivariate Analysis (area utama)
5. Social Network Analysis
6. Statistical Process Control.

Email: maman_djauhari@yahoo.com.

2. Dyah Erny Herwindiati

Dyah Erny Herwindiati adalah Guru Besar Teknik Informatika Universitas Tarumanagara, sejak tahun 2014. Dia menyelesaikan program doktoralnya di Departemen Matematika Institut Teknologi Bandung (ITB) pada tahun 2006, dengan Disertasi berjudul *A New Criterion of Robust Estimatimation for Location and Covariance Matrix and its Application in Outlier Labeling*. Saat ini Dyah mengabdikan diri pada kepemimpinan akademis sebagai dekan Fakultas Teknologi Informasi di Universitas Tarumanagara, Jakarta, periode 2014 – 2026.

Sejak meraih gelar doktornya, dia telah berada di garis depan dalam mengembangkan algoritma inovatif dan robust yang dirancang khusus untuk aplikasi pengolahan data citra satelit (*remote sensing data*). Algoritma tersebut telah meningkatkan akurasi dan keandalan analisis data di berbagai bidang.

Dia juga pernah menjadi *Visiting Researcher* di *Mathematics Department*, Conservatoire National des Arts et Métiers (CNAM) University, Prancis, dan di *Mathematics Department*, Universiti Teknologi Malaysia (UTM), Malaysia.

Pada tahun 2014, keahlian dan kontribusinya di bidang *Data Mining* diakui negara ketika Menteri Pendidikan Indonesia secara resmi melantiknya sebagai Guru Besar. Penelitiannya telah memberikan dampak signifikan pada berbabai bidang, khususnya bidang *Machine*

Learning dan *Data Science*. Kontribusinya di dunia akademis telah memperkaya pengalaman belajar dan mendorong kolaborasi yang menjembatani penelitian teoretis dan aplikasi praktis di bidang informatika.

Riset yang ditekuni meliputi *Data Mining, Machine Learning, Remote Sensing,* dan *Robust Statistics.*

Dyah dapat dihubungi di Email: dyahh@fti.untar.ac.id

Call/WA: +62 8888 03 11 30
Website: www.campustaka.com

KOMPUTER

ISBN 978-634-7639-7-6

9 786349 639576