ISBN 978-623-178-228-5 (no.jil.lengkap)
ISBN 978-623-125-611-9 (jil.2)

PENGANTAR METODE STATISTIKA DALAM ILMU METEOROLOGI DAN KLIMATOLOGI

Jilid 2

PENULIS:

- Nurtiti Sunusi
- Erna Tri Herdiani
- Siswanto
- Andi Kresna Jaya
- Muhammad Fadil

PENGANTAR METODE STATISTIKA

DALAM ILMU METEOROLOGI DAN KLIMATOLOGI

Jilid 2

Penulis

Nurtiti Sunusi
Erna Tri Herdiani
Siswanto
Andi Kresna Jaya
Muhammad Fadil

GETPRESS INDONESIA

PENGANTAR
METODE STATISTIKA
DALAM ILMU METEOROLOGI DAN KLIMATOLOGI
Jilid 2

Penulis : Nurtiti Sunusi
Erna Tri Herdiani
Siswanto
Andi Kresna Jaya
Muhammad Fadil

ISBN : 978-623-125-611-9
(jil.2)

Editor : Mila Sari, M.Si.
Desain Sampul dan Tata Letak : Tri Putri Wahyuni, S.Pd.

PENERBIT GET PRESS INDONESIA
Anggota IKAPI No. 033/SBA/2022
Jl. Palarik RT 01 RW 06 Kelurahan Air Pacah
Kecamatan Koto Tangah Padang Sumatera Barat

website: www.getpress.co.id
email: adm.getpress@gmail.com

Cetakan pertama, Februari 2025

KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Kuasa, yang telah memberikan rahmat-Nya sehingga buku Pengantar Metode Statistika dalam Ilmu Meteorologi dan Klimatologi ini dapat diselesaikan dengan sebaik-baiknya.

Buku ini memberikan informasi secara lengkap mengenai materi apa saja yang akan mereka pelajari yang berasal dari berbagai sumber terpercaya yang berguna sebagai tambahan wawasan mengenai bab-bab yang dipelajari tersebut.

Penyusun meyakini bahwa dalam pembuatan buku Pengantar Metode Statistika dalam Ilmu Meteorologi dan Klimatologi ini masih jauh dari sempurna. Oleh karena itu penyusun mengharapkan kritik dan saran yang membangun guna penyempurnaan modul praktikum ini dimasa yang akan datang.

Akhir kata, penyusun mengucapkan banyak terima kasih kepada semua pihak yang telah membantu baik secara langsung maupun tidak langsung.

Makassar, September 2024

Penyusun

DAFTAR ISI

kata pengantar ..... **I**
DAFTAR ISI ..... **II**
DAFTAR GAMBAR ..... **V**
DAFTAR TABEL ..... **XVI**
BAB 1 PENGANTAR **1**
1.1 Robustness **1**
1.2 Resistance **2**
1.3 Quantiles **3**
1.3.1 Definisi Quantiles 3
1.3.2 Penghitungan Quantiles 4
1.3.3 Contoh Quantiles Umum 4
1.4 Analisis Data Eksploratif (Exploratory Data Analysis - EDA) **5**
1.4.1 Boxplot 5
1.4.2 Scatterplot 7
1.4.3 Autokorelasi 9
BAB 2 DISTRIBUSI EMPIRIS DAN ANALISIS DATA EKSPLORASI **14**
2.1 Penskalaan Ulang (*Rescalling*) **14**
2.1.1 Transformasi 15
2.1.2 Standarisasi Anomali (Normalisasi) 25
2.2 Teknik Eksplorasi Data Berpasangan **29**
2.2.1 Scatterplot 30
2.2.2 Korelasi Pearson 32
2.2.3 Korelasi Spearman Rank dan Kendall τ 45
2.2.4 Fungsi Autokorelasi 47
2.2.5 Fungsi Autokorelasi 50
2.3 Teknik Eksplorasi Data Dimensi Tinggi **52**
2.3.1 Plot Bintang 53
2.3.2 Glyph Scatterplot 55
2.3.3 The Rotating Scatterplot 59

2.3.4 Matriks Korelasi 62
2.3.5 Matriks Scatterplot 66
2.3.6 Peta Korelasi 69
Latihan 75
BAB 3 DISTRIBUSI PROBABILITAS PARAMETRIK 76
3.1 Pendahuluan 76
3.1.1 Distribusi parametrik vs. Empiris 76
3.1.2 Apa itu Distribusi Parametrik? 78
3.1.3 Parameter vs Statistik 79
3.1.4 Distribusi Diskrit vs. Kontinu 80
3.2 Distribusi Diskrit 81
3.2.1 Distribusi Binomial 81
3.2.2 Distribusi Geometrik 90
3.2.3 Distribusi Binomial Negatif 91
3.2.4 Distribusi Poisson 98
3.3 Ekspektasi Statistik 103
3.3.1 Nilai Harapan dari Variabel Acak 103
3.3.2 Nilai yang diharapkan dari Fungsi Variabel Acak 105
3.4 Distribusi Kontinyu 108
3.4.1 Fungsi Distribusi dan Nilai Harapan 109
3.4.2 Distribusi Gaussian 114
3.4.3 Distribusi Gamma 132
3.4.4 Distribusi Beta 148
3.4.5 Distribusi Nilai Ekstrim 151
3.4.6 Distribusi campuran 163
3.5 Penilaian Kualitatif tentang Goodness of Fit 169
3.5.1 Superposisi Distribusi Parametrik Terpasang dan Histogram Data 169
3.5.2 Diagram Kuantil-Kuantil (Q–Q) 173
3.6 Pemasangan Parameter Kemungkinan Maksimum 176
3.6.1 Fungsi kemungkinan (Likelihood) 176
3.6.2 Metode Newton-Raphson 180
3.6.3 Algoritma EM 182
3.6.4 Distribusi Sampling Estimasi Kemungkinan Maksimum 189

3.7 Simulasi Statistik 191
3.7.1 Generator Angka Acak Seragam 192
3.7.2 Pembuatan bilangan acak tidak seragam dengan inversi 197
3.7.3 Pembangkitan Bilangan Acak Tidak Seragam dengan Penolakan 200
3.7.4 Metode Box-Muller untuk membangkitkan bilangan acak Gaussian 204
3.7.5 Simulasi Distribusi Campuran Dan Perkiraan Densitas Kernel 206
Latihan 210
DAFTAR PUSTAKA 214
LAMPIRAN 220
Tabel A 220
Kumpulan Contoh Data 220
Tabel B 224
Tabel Probabilitas 224

DAFTAR GAMBAR

Gambar 1.1 Boxplot

Gambar 1.2 Scatterplot

Gambar 2.1 Grafik transformasi pangkat pada Persamaan 1.2 untuk nilai parameter transformasi λ. Untuk $\lambda = 1$ transformasinya linier dan tidak menghasilkan perubahan bentuk data. Untuk $\lambda < 1$ transformasi mengurangi semua nilai data dan nilai yang lebih besar akan lebih terpengaruh. Efek sebaliknya dihasilkan oleh transformasi dengan $\lambda > 1$.

Gambar 2.2 Pengaruh transformasi pangkat dengan $\lambda < 1$ pada kumpulan data dengan kemiringan positif (kurva tebal). Panah menunjukkan bahwa transformasi memindahkan semua titik ke kiri, dengan nilai yang lebih besar dipindahkan lebih banyak. Distribusi yang dihasilkan (kurva tipis) cukup simetris.

Gambar 2.3 Pengaruh transformasi pangkat dalam Persamaan 1.2 pada data curah hujan total bulan Januari untuk Ithaca, 1933–1982 (Tabel A.2). Data asli ($\lambda = 1$) sangat miring ke kanan, dengan nilai terbesar berada jauh di luar. Transformasi akar kuadrat ($\lambda = 0.5$) sedikit meningkatkan simetri. Transformasi logaritmik ($\lambda = 0$) menghasilkan distribusi yang cukup simetris. Ketika mengalami

transformasi akar kuadrat terbalik yang lebih ekstrem ($\lambda = -0.5$) data mulai menunjukkan kemiringan negatif. Transformasi logaritmik akan dipilih sebagai yang terbaik berdasarkan statistik Hinkley $d\lambda$ (Persamaan 1.3), dan log-likelihood Gaussian (Persamaan 1.4).

Gambar 2.4 Standarisasi selisih antar standarisasi anomali tekanan permukaan laut bulanan di Tahiti dan Darwin (Indeks Osilasi Selatan), 1960–2002. Nilai bulanan individu telah dihaluskan dalam waktu.

Gambar 2.5 Scatterplot untuk suhu maksimum dan minimum harian selama Januari 1987 di Ithaca, New York. Lingkaran tertutup mewakili hari dengan curah hujan setidaknya 0.01 inci (setara cairan).

Gambar 2.6 Awan titik hipotetis dalam dua dimensi yang menggambarkan mekanisme koefisien korelasi Pearson (Persamaan 1.7). Dua sampel berarti membagi bidang menjadi empat kuadran, diberi nomor I–IV.

Gambar 2.7 Scatterplot dari dua set data berpasangan buatan pada Tabel 1.4. Korelasi Pearson untuk data pada panel (a) (Set I pada Tabel 1.4) hanya sebesar 0,88 mewakili kekuatan hubungan yang kurang terwakili, menunjukkan bahwa ukuran korelasi ini

tidak kuat. Korelasi Pearson untuk data di panel (b) (Teorema II) adalah 0,61, yang mencerminkan pengaruh luar biasa dari titik eksterior tunggal dan menggambarkan tidak adanya perlawanan.

Gambar 2.8 Konstruksi rangkaian waktu tergeser dari data suhu maksimum Ithaka untuk Januari 1987. Pergeseran data satu hari menyisakan 30 pasangan data (terlampir dalam kotak) untuk menghitung koefisien autokorelasi lag 1.

Gambar 2.9 Contoh fungsi autokorelasi untuk data suhu maksimum dari Ithaca pada Januari 1987. Korelasinya adalah 1 untuk k = 0 karena data sesaat berkorelasi sempurna dengan diri mereka sendiri. Fungsi autokorelasi turun menjadi hampir nol untuk $k \geq 10$.

Gambar 2.10 Bagan bintang selama lima hari terakhir dalam data Januari 1987 pada Tabel A.1, dengan sumbu berlabel hanya untuk bintang 27 Januari. Perkiraan simetri radial dalam bagan ini mencerminkan korelasi antara variabel serupa di dua lokasi, dan perluasan bintang dari waktu ke waktu menunjukkan hari yang lebih hangat dan lebih basah menjelang akhir bulan.

Gambar 2.11 Glyph scatterplot distribusi frekuensi bivariat dari perkiraan dan suhu maksimum

musim dingin harian yang diamati untuk Minneapolis, 1980–1981 hingga 1985–1986. Suhu telah dibulatkan ke interval 5°F, dan mesin terbang melingkar telah diskalakan untuk memiliki area yang sebanding dengan hitungan (inset).

Gambar 2.12 Histogram bivariat disajikan dalam tampilan perspektif dari data yang sama disajikan sebagai scatterplot mesin terbang pada **Gambar 2.11**. Meskipun titik data pada bidang prediksi-pengamatan terletak di ekor vertikal dan titik pada diagonal 1:1 selanjutnya ditunjukkan oleh lingkaran terbuka, proyeksi dari tiga ke dua dimensi memperumit interpretasi gambar. Dari Murphy dkk. (1989).

Gambar 2.13 Mesin terbang canggih yang dikenal sebagai model stasiun meteorologi yang mewakili tujuh besaran secara bersamaan. Saat diplot pada peta, itu juga menambahkan dua variabel lokasi (lintang dan bujur), meningkatkan dimensi plot menjadi sembilan magnitudo, yang setara dengan sebaran data cuaca mesin terbang.

Gambar 2.14 Empat potret evolusi plot rotasi tiga dimensi dari data Guayaquil Juni pada Tabel A.3, menunjukkan lima tahun El Niño sebagai lingkaran. Sumbu suhu tegak lurus terhadap

halaman dan memanjang keluar halaman pada panel (a), dan tiga panel berikutnya menunjukkan perubahan perspektif ketika sumbu suhu diputar pada bidang halaman, ke bawah dan ke kiri. Ilusi visual awan titik yang mengambang di ruang tiga dimensi jauh lebih besar dalam tampilan langsung yang bergerak terus menerus.

Gambar 2.15 Tata letak matriks korelasi, [R]. Korelasi ri, j antara semua pasangan variabel yang mungkin disusun sedemikian rupa sehingga indeks pertama i menunjukkan nomor baris dan indeks kedua j menunjukkan nomor kolom.

Gambar 2.16 Matriks Scatterplot untuk data Januari 1987.

Gambar 2.17 Peta korelasi satu titik tekanan permukaan tahunan di lokasi di seluruh dunia dengan yang ada di Jakarta, Indonesia. Korelasi negatif yang kuat dari -0.8 di Pulau Paskah terkait dengan fenomena El Nino-Osilasi Selatan. Dari Bjerknes (1969).

Gambar 2.18 Telekonektivitas atau nilai absolut dari korelasi negatif terkuat dari masing-masing peta korelasi satu titik yang banyak digambar pada titik grid dasar, untuk ketinggian 500 mb di musim dingin. Dari Wallace dan Blackmon (1983).

Gambar 3.1 Fungsi distribusi probabilitas (a) dan fungsi distribusi probabilitas kumulatif (b) waktu tunggu $x + k$ tahun agar Danau Cayuga membeku k kali, menggunakan distribusi binomial negatif pada Persamaan 2.6.

Gambar 3.2 Histogram jumlah tornado yang dilaporkan setiap tahun di Negara Bagian New York selama 1959–1988 (garis putus-putus) dan distribusi Poisson yang sesuai dengan $\mu = 4.6$ tornado/tahun (garis tebal hitam).

Gambar 3.3 Fungsi kepadatan probabilitas hipotetis f(x) untuk variabel acak nonnegatif X. Mengevaluasi f(x) dengan sendirinya tidak berarti dalam hal probabilitas untuk nilai X tertentu. Probabilitas ditemukan dengan mengintegrasikan bagian-bagian dari f(x).

Gambar 3.4 Fungsi kepadatan probabilitas untuk distribusi Gaussian. Rata-rata μ, menempatkan pusat distribusi simetris ini, dan standar deviasi, σ, mengontrol sejauh mana distribusi menyebar. Hampir semua probabilitas berada dalam $\pm 3\sigma$ dari rata-rata.

Gambar 3.5 Pandangan perspektif distribusi normal bivariat dengan $\sigma x = \sigma y$ dan $= 0.75$. Garis individu yang mewakili punuk distribusi bivariat berbentuk distribusi Gaussian (univariat), menunjukkan bahwa distribusi

bersyarat dari x memiliki beberapa nilai y adalah Gaussian sendiri.

Gambar 3.6 Distribusi Gaussian menggambarkan distribusi tak bersyarat untuk suhu maksimum harian Januari di Canandaigua dan distribusi bersyarat dengan asumsi bahwa suhu maksimum di Ithaca adalah 25°F. Korelasi yang tinggi antara suhu maksimum di kedua lokasi berarti bahwa distribusi bersyarat jauh lebih tajam, mencerminkan ketidakpastian yang jauh berkurang.

Gambar 3.7 Fungsi densitas distribusi gamma untuk empat nilai parameter bentuk α

Gambar 3.8 Curah hujan total untuk Januari 1989 di atas Amerika Serikat yang bersebelahan dinyatakan sebagai nilai persentil dari distribusi gamma lokal. Bagian timur dan barat lebih kering dari biasanya, dan bagian tengah negara itu lebih basah. Oleh Arkin (1989).

Gambar 3.9 Parameter bentuk distribusi gamma untuk curah hujan Januari di Amerika Serikat. Distribusi di barat daya sangat miring, dan untuk sebagian besar lokasi di timur jauh lebih simetris. Distribusi dilengkapi dengan data dari 30 tahun 1951-1980. Dari Wilks dan Eggleston (1992).

Gambar 3.10 Representasi kategori presipitasi pada **Gambar 3.8** dalam bentuk fungsi densitas distribusi gamma dengan α = 2. Hasil yang lebih kering dari persentil ke-10 berada di sebelah kiri $q0.1$. Area dengan curah hujan antara persentil ke-30 dan ke-70 (antara $q0.3$ dan $q0.7$) tidak akan diarsir pada peta. Curah hujan di 10% terbasah dari distribusi klimatologi terletak di sebelah kanan $q0.9$.

Gambar 3.11 Lima contoh fungsi kerapatan probabilitas untuk distribusi beta. Gambar cermin dari distribusi ini diperoleh dengan membalik parameter p dan q.

Gambar 3.12 Fungsi kepadatan probabilitas distribusi Weibull untuk empat nilai parameter bentuk a.

Gambar 3.13 Fungsi kerapatan peluang untuk campuran (Persamaan 2.63) dari dua distribusi Gaussian cocok dengan data suhu Guayaquil bulan Juni (Tabel A3). Hasilnya sangat mirip dengan perkiraan densitas kernel yang diperoleh dari data yang sama, Gambar 1.8b

Gambar 3.14 Plot kontur PDF dari distribusi campuran Gaussian bivariat yang dipasang pada ansambel 51 prakiraan untuk suhu 2 m dan kecepatan angin 10 m, dibuat dengan waktu tunggu 180 jam. Kecepatan angin pertama kali ditransformasikan akar kuadrat untuk

membuat distribusi univariatnya lebih Gaussian. Titik menandai prakiraan individu dari 51 anggota ansambel. Dua densitas Gaussian bivariat konstituen $f1x$ dan $f2x$ masing-masing berpusat pada 1 dan 2, dan garis halus menunjukkan kurva level dari campurannya fx yang dibentuk dengan w = 0.57. Interval kontur tetap adalah 0,05 , dan garis putus-putus berat dan ringan masing-masing adalah 0,01 dan 0,001. Diadaptasi dari Wilks (2002b).

Gambar 3.15 Histogram data curah hujan Januari 1933–1982 di Ithaca dari Tabel A2 dengan PDF gamma (padat) dan Gaussian (rusak). Masing-masing dari dua fungsi kerapatan telah dikalikan dengan A = 25 karena lebar bin adalah 0,5 inci dan terdapat 50 pengamatan. Rupanya distribusi gamma memberikan representasi data yang masuk akal. Distribusi Gaussian kurang mewakili ekor kanan dan menyiratkan probabilitas presipitasi negatif yang tidak nol.

Gambar 3.16 Plot kuantil-kuantil untuk gamma (o) dan Gaussian (x) sesuai dengan curah hujan Januari 1933–1982 di Ithaca pada Tabel A2. Jumlah curah hujan yang diamati berada pada posisi vertikal dan jumlah yang berasal dari distribusi yang dipasang menggunakan

posisi plot Tukey berada pada posisi horizontal. Garis diagonal menunjukkan korespondensi 1:1.

Gambar 3.17 1000 pasangan variabel acak seragam yang tidak tumpang tindih di bagian kecil persegi yang didefinisikan oleh $0 < un < 1$ dan $0 < un + 1 < 1$; dihasilkan menggunakan Persamaan 2.79, dan $a = 16807$, $c = 0$ dan $M = 231 - 1$. Wilayah kecil ini berisi 17 dari sekitar 65.000 garis sejajar dimana pasangan berurutan jatuh di seluruh unit persegi.

Gambar 3.18 Ilustrasi pembangkitan variabel eksponensial dengan inversi. Kurva halus adalah CDF (Persamaan 2.46) dengan rata-rata $\beta = 2.7$. Variabel seragam $u = 0.9473$ diubah menjadi variabel eksponensial yang dihasilkan $x = 7.9465$ dengan kebalikan dari CDF. Gambar ini juga mengilustrasikan bahwa inversi menghasilkan transformasi monoton dari varian seragam yang mendasarinya.

Gambar 3.19 Representasi simulasi dari kerapatan kuartik (biweight), $fx = 15161 - x22$ (Tabel 1.1), menggunakan kerapatan segitiga (Tabel 1.1) sebagai amplop, dengan $c = 1.12$. Dua puluh lima kandidat x disimulasikan dari kerapatan segitiga, 21 di antaranya diterima ( ) karena mereka juga berada di bawah distribusi fx yang akan disimulasikan, dan 4 ditolak (O)

karena berada di luarnya. Garis abu-abu terang menunjukkan nilai yang disimulasikan pada sumbu horizontal.

DAFTAR TABEL

Tabel 1.1 Curah hujan Ithaca Januari 1933–1982 dari Tabel A.2 ($\lambda = 1$). Data telah disortir dengan transformasi pangkat pada Persamaan 1.2 diterapkan untuk $\lambda = 1$, $\lambda = 0.5$, $\lambda = 0$, dan $\lambda = -0.5$. Plot skematis dari data ini ditunjukkan pada **Gambar 2.3**.

Tabel 1.2 Set data berpasangan artifisial untuk contoh korelasi.

Tabel 1.3 Matriks korelasi untuk data pada Tabel A.1. Hanya segitiga bawah dari matriks yang ditunjukkan untuk menghilangkan redudansi dan nilai diagonal yang tidak jelas. Matriks kiri berisi korelasi momen produk Pearson dan matriks kanan berisi korelasi peringkat Spearman.

Tabel 2.1 Data tahunan Danau Cayuga saat membeku.

Tabel 2.2 Nilai fungsi Gamma, Γk (Persamaan 2.7), untuk $1.00 \leq k \leq 1.99$.

Tabel 2.3 Jumlah Tornado Setiap Tahun di Negara Bagian New York, 1959–1988.

Tabel 2.4 Nilai yang diharapkan (rata-rata) dan varians untuk empat probabilitas diskrit fungsi distribusi yang dijelaskan dalam Bagian 4.2 berkenaan dengan parameter distribusinya.

Tabel 2.5 Probabilitas binomial untuk N = 3 dan p = 0.5 dan konstruksi ekspektasi E[X] dan E[X2] sebagai rata-rata pembobotan probabilitas.

Tabel 2.6 Mencantumkan rata-rata dan varian untuk distribusi yang akan dijelaskan di bagian ini terkait dengan parameter distribusi
Tabel 2.7 Jumlah curah hujan harian maksimum tahunan (inci) di Charleston, Carolina Selatan, 1951–1970
Tabel 2.8 Kemajuan algoritme EM selama tujuh iterasi yang diperlukan agar sesuai dengan campuran PDF Gaussian yang ditunjukkan pada **Gambar 3.13**

BAB 1
PENGANTAR

Dalam statistik, konsep **Robustness** dan **Resistance** sangat penting untuk menangani data yang mungkin mengandung kesalahan atau ketidaksesuaian dengan asumsi distribusi tertentu. Kedua sifat ini membantu mengurangi pengaruh data yang tidak diinginkan, seperti pencilan atau distribusi yang tidak mengikuti asumsi normalitas. Dalam meteorologi dan klimatologi, data observasi atmosferik dan model numerik menghasilkan sejumlah besar data numerik. Salah satu tujuan utama analisis data ini adalah untuk memahami distribusi dan pola data melalui distribusi empiris dan teknik eksplorasi data.

1.1 Robustness

Suatu metode statistik dikatakan **robust** jika performanya tidak terlalu terpengaruh oleh pelanggaran asumsi distribusi tertentu. Sebagai contoh, rata-rata (*mean*) merupakan estimator terbaik untuk pusat distribusi pada data normal. Namun, jika data mengandung pencilan atau memiliki distribusi yang berat di ekor, rata-rata bisa memberikan hasil yang sangat menyimpang dari nilai pusat data yang sebenarnya.

Contoh Robustness: Ketika menggunakan sampel dari distribusi normal $X \sim N(\mu, \sigma^2)$, rata-rata sampel $\bar{X}$ adalah estimator yang optimal jika distribusi benar-benar normal. Namun, jika data mengandung pencilan yang kuat, $\bar{X}$ mungkin menjadi tidak representatif. Sebagai alternatif, *median* lebih robust terhadap pencilan dibandingkan rata-rata karena kurang terpengaruh oleh nilai ekstrem dalam data.

1.2 Resistance

Sebuah statistik disebut **resistant** jika hasil estimasinya tidak terpengaruh oleh perubahan ekstrem dalam beberapa nilai data. Sifat resistant berfokus pada bagaimana statistik tetap stabil walaupun ada data pencilan atau variasi besar dalam sebagian kecil data.

Contoh Resistance: Interquartile Range (IQR) adalah ukuran sebaran yang resistant karena hanya memperhitungkan 50% data di tengah. Sebagai contoh:

$$IQR = Q_3 - Q_1$$

di mana Q_3 adalah kuartil ketiga dan Q_1 adalah kuartil pertama. Karena IQR hanya bergantung pada data tengah, ia tidak terpengaruh oleh pencilan atau perubahan ekstrem dalam nilai minimum atau maksimum data.

Catatan

Robustness dan resistance adalah konsep yang membantu dalam pemilihan metode statistik yang sesuai untuk data atmosferik dan klimatologis, di mana kehadiran pencilan atau pelanggaran asumsi normalitas mungkin terjadi. Penggunaan statistik yang robust dan resistant dapat menghasilkan analisis yang lebih reliabel dalam situasi di mana data tidak sepenuhnya memenuhi asumsi klasik statistik.

1.3 Quantiles

Dalam analisis data statistik, terutama dalam meteorologi dan klimatologi, **quantiles** (atau *fractiles*) adalah ukuran yang digunakan untuk membagi data menjadi beberapa bagian berdasarkan proporsi tertentu. Quantile menunjukkan nilai-nilai data pada persentase tertentu, misalnya median adalah quantile yang membagi data menjadi dua bagian yang sama besar.

1.3.1 Definisi Quantiles

Secara formal, quantile q_p adalah suatu nilai dalam kumpulan data yang melebihi proporsi data sebesar p, dengan $0 \leq p \leq 1$. Artinya, jika $p = 0.25$, maka $q = 0.25$ atau kuartil pertama (Q_1) adalah nilai yang memisahkan 25% data terendah dari 75% data lainnya.

Secara umum, quantile q_p didefinisikan sebagai berikut:

$$q_p = \text{nilai yang melebihi } p \times 100\% \text{ dari data}$$

1.3.2 Penghitungan Quantiles

Untuk menghitung quantile pada sampel data, langkah pertama adalah mengurutkan data dari nilai terendah hingga tertinggi. Data yang telah diurutkan ini disebut sebagai *order statistics*. Misalkan terdapat data $x_1, x_2, \cdots, x_n$, maka order statistics disimbolkan sebagai $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$, di mana $x_{(i)}$ adalah data ke-i terkecil.

1.3.3 Contoh Quantiles Umum

Beberapa quantiles yang sering digunakan meliputi:

- **Median** ($q_{0.5}$): membagi data menjadi dua bagian yang sama, yaitu 50% data di bawah dan 50% data di atas nilai median.
- **Kuartil Pertama** ($q_{0.25}$): memisahkan 25% data terendah dari 75% data tertinggi.
- **Kuartil Ketiga** ($q_{0.75}$): memisahkan 75% data terendah dari 25% data tertinggi.

Rumus untuk menghitung median pada kumpulan data dengan n pengamatan:

$$q_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{jika } n \text{ ganjil} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+1}{2}\right)}}{2} & \text{jika } n \text{ genap} \end{cases}$$

Contoh Perhitungan Quantiles Misalkan terdapat data suhu harian $x = \{10, 12, 15, 20, 22, 25, 30\}$.

- **Median** ($q_{0.5}$): Karena terdapat 7 data (ganjil), median adalah data ke-4 yaitu $x_{(4)} = 20$.
- **Kuartil Pertama** ($q_{0.25}$): Kuartil pertama adalah data posisi $0.25 \times (7 + 1) = 2$. Jadi $q_{0.25} = x_{(2)} = 12$.
- **Kuartil Ketiga** ($q_{0.75}$): Kuartil ketiga adalah data posisi $0.75 \times (7 + 1) = 6$. Jadi $q_{0.75} = x_{(6)} = 25$.

1.4 Analisis Data Eksploratif (Exploratory Data Analysis - EDA)

EDA bertujuan untuk mengeksplorasi data secara visual untuk menemukan pola, anomali, dan informasi penting lainnya. Beberapa alat utama dalam EDA meliputi:

1.4.1 Boxplot

Boxplot adalah grafik yang menunjukkan distribusi data berdasarkan lima ukuran ringkasan: minimum, kuartil pertama (Q_1), median, kuartil ketiga (Q_3), dan maksimum. Dalam klimatologi, boxplot dapat digunakan untuk menggambarkan rentang suhu atau kelembaban selama satu bulan atau musim tertentu.

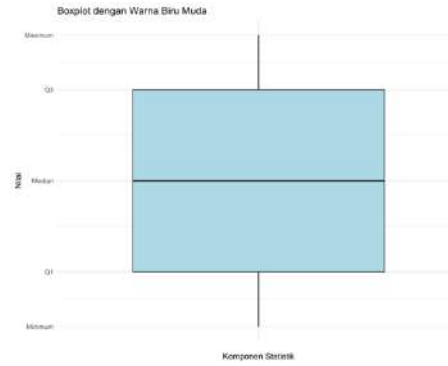
Contoh code R:

```
# Memuat library yang diperlukan
library(ggplot2)

# Contoh data yang merepresentasikan boxplot
data <- data.frame(
  Nilai = c(10, 25, 50, 75, 90)
)

# Membuat boxplot dengan ggplot2
ggplot(data, aes(x = "", y = Nilai)) +
  geom_boxplot(
    fill = "lightblue", # Warna biru muda
    outlier.shape = NA  # Menyembunyikan outlier
  ) +
  scale_y_continuous(
    breaks = c(10, 25, 50, 75, 90),
    labels   =   c("Minimum",   "Q1",   "Median",   "Q3",
"Maximum")
  ) +
  labs(
    x = "Komponen Statistik",
    y = "Nilai",
    title = "Boxplot dengan Warna Biru Muda"
  ) +
  theme_minimal()
```

Output:



Gambar 1.1 Boxplot

1.4.2 Scatterplot

Scatterplot memvisualisasikan hubungan antara dua variabel, seperti suhu dan kelembaban. Ini membantu untuk mendeteksi korelasi atau pola antara variabel yang diukur pada waktu atau tempat yang sama.

Scatterplot adalah grafik yang menunjukkan hubungan antara dua variabel numerik melalui titik-titik yang terdistribusi di dalam bidang kartesian. Setiap titik dalam scatterplot mewakili satu pengamatan dalam data, dengan posisi horizontal (sumbu x) dan vertikal (sumbu y) sesuai dengan nilai dari dua variabel yang berbeda.

Scatterplot berguna dalam melihat korelasi antara dua variabel, mengidentifikasi pola, dan mendeteksi adanya pencilan. Misalnya, dalam meteorologi, scatterplot dapat

digunakan untuk memvisualisasikan hubungan antara suhu dan kelembaban.

Contoh Misalkan kita memiliki data mengenai suhu harian dan kelembaban relatif selama beberapa hari. Dengan menggunakan scatterplot, kita dapat memvisualisasikan apakah ada korelasi antara suhu dan kelembaban.

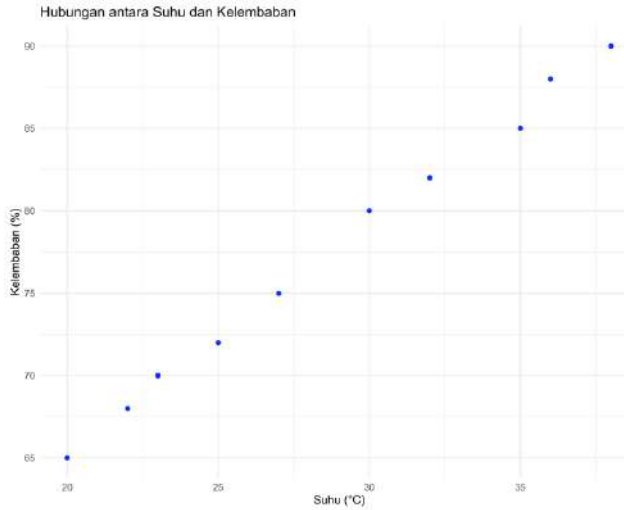
Code R:

```
# Memuat library yang diperlukan
library('ggplot2')

# Membuat data frame untuk plot
data <- data.frame(
  Suhu = c(20, 22, 23, 25, 27, 30, 32, 35, 36, 38),
  Kelembaban = c(65, 68, 70, 72, 75, 80, 82, 85, 88,
90)
)

# Membuat scatter plot dengan ggplot2
scatterplt <- ggplot(data, aes(x = Suhu, y =
Kelembaban)) +
  geom_point(shape = 16, color = "blue", size=2) + #
Titik dengan bentuk bulat dan warna biru
  labs(
    title = "Hubungan antara Suhu dan Kelembaban",
    x = "Suhu (°C)",
    y = "Kelembaban (%)"
  ) +
  theme_minimal() # Tema minimalis agar tampilan rapi
```

Output:



Gambar 1.2. Scatterplot

Interpretasi pada scatterplot di atas, kita dapat melihat bahwa titik-titik cenderung meningkat ke arah kanan atas, yang menunjukkan adanya korelasi positif antara suhu dan kelembaban. Semakin tinggi suhu, kelembaban cenderung lebih tinggi pula.

1.4.3 Autokorelasi

Pengertian Autokorelasi

Autokorelasi adalah ukuran yang menunjukkan seberapa kuat keterkaitan suatu data dengan nilai-nilai sebelumnya dalam suatu deret waktu. Dalam statistik, autokorelasi digunakan untuk menganalisis pola berulang dalam data yang

dikumpulkan secara berurutan dalam waktu, seperti data cuaca harian, harga saham, atau hasil pengukuran atmosfer.

Autokorelasi digunakan untuk mengukur keterkaitan antar data dalam waktu berbeda, yang penting untuk melihat pola musiman. Dalam meteorologi, fungsi autokorelasi dapat mengidentifikasi ketergantungan temporal pada data iklim harian atau bulanan, misalnya:

$$r_k = \frac{\sum_{i=1}^{n-k}(X_i - \bar{X})(X_{i+k} - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

di mana X_i adalah nilai data pada waktu ke-i, k adalah lag, dan $\bar{X}$ adalah rata-rata dari data tersebut.

Autokorelasi pada lag k, atau jarak waktu k, mengukur hubungan antara data pada waktu t dan data pada waktu $t - k$. Autokorelasi didefinisikan sebagai:

$$r_k = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

di mana:
X_t adalah nilai data pada waktu t
$\bar{X}$ adalah rata-rata dari data
n adalah jumlah data dalam deret waktu
k adalah lag atau pergeseran waktu

Autokorelasi sangat berguna dalam meteorologi untuk mengidentifikasi pola musiman, tren, atau siklus dalam data cuaca, seperti suhu atau curah hujan.

Contoh Autokorelasi

Misalkan kita memiliki data suhu harian selama 10 hari: $\{20, 21, 20, 22, 24, 23, 25, 24, 26, 25\}$. Kita ingin menghitung autokorelasi pada lag 1 untuk melihat seberapa terkait suhu hari ini dengan suhu hari sebelumnya.

1. Hitung rata-rata $\bar{X}$ dari data

$$\bar{X} = \frac{20 + 21 + 20 + 22 + 24 + 23 + 25 + 24 + 26 + 25}{10} = 23$$

2. Gunakan rumus autokorelasi untuk lag 1

$$r_k = \frac{\sum_{t=1}^{9}(X_t - 23)(X_{t+k} - 23)}{\sum_{t=1}^{10}(X_t - 23)^2}$$

3. Hasil perhitungan ini memberikan nilai autokorelasi untuk lag 1

Catatan jika k memiliki nilai yang tinggi untuk lag tertentu, ini menunjukkan bahwa data tersebut memiliki pola berulang atau siklus pada interval waktu tersebut. Dalam contoh ini, jika r_1 tinggi, berarti suhu hari ini sangat terkait dengan suhu hari sebelumnya, yang menunjukkan pola berkelanjutan dalam deret waktu.

Distribusi empiris dan teknik EDA memberikan pemahaman awal tentang data meteorologi atau klimatologi,

memungkinkan kita untuk mendeteksi pola, anomali, dan karakteristik penting lainnya. Langkah awal ini penting dalam analisis data atmosferik sebelum melangkah ke tahap analisis inferensial atau pemodelan statistik lanjutan.

Code R:

```
## data hari ##
data_hari <- c(20,21,20,22,24,23,25,24,26,25)

## Function Autokorelasi ##
autokorelasi <- function(data, lag){
  x <- as.vector(data)
  k <- lag
  n <- length(x)

  tmp0 <- 0
  for (t in 1:(n-k)) {
    tmp0 <- tmp0 + ((x[t]-mean(x)) * (x[t+k]-mean(x)))
  }

  hasil <- tmp0/sum((x-mean(x))^2)

  message('Autokorelasi: ', round(hasil,6), ' untuk lag
',k)

  return(hasil)
}

autokorelasi_data <- autokorelasi(data_hari, lag = 1)
```

Output:

```
Autokorelasi: 0.595238 untuk lag 1
```

BAB 2
DISTRIBUSI EMPIRIS DAN ANALISIS DATA EKSPLORASI

2.1 Penskalaan Ulang (*Rescalling*)

Rescaling atau penskalaan ulang adalah proses mengubah skala atau rentang nilai data untuk memenuhi kebutuhan tertentu, seperti meningkatkan performa algoritma pembelajaran mesin atau menyederhanakan interpretasi data. Rescaling sering digunakan dalam analisis statistik dan pembelajaran mesin karena banyak algoritma memerlukan data dalam rentang tertentu untuk berfungsi secara optimal. Ada beberapa metode yang umum digunakan untuk rescaling data:

Ada kemungkinan bahwa skala pengukuran asli dapat mengaburkan fitur-fitur penting dalam sekumpulan data. Jika demikian, analisis dapat difasilitasi, atau dapat menghasilkan hasil yang lebih terbuka jika data terlebih dahulu mengalami transformasi matematis. Transformasi tersebut juga dapat sangat berguna untuk membantu data atmosfer sesuai dengan asumsi analisis regresi atau memungkinkan penerapan metode statistik multivariat yang dapat mengasumsikan distribusi Gaussian. Dalam terminologi

analisis data eksplorasi, transformasi data tersebut dikenal sebagai ekspresi (penskalaan) ulang pada data.

2.1.1 Transformasi

Seringkali, data diubah untuk membuat distribusi nilai lebih simetris, dan hasilnya memungkinkan penggunaan teknik statistik yang lebih umum dan tradisional. Terkadang, transformasi penghasil simetri dapat membuka analisis eksplorasi, seperti yang dijelaskan dalam bab ini. Dengan meluruskan hubungan antara dua variabel, transformasi ini dapat membantu dalam membandingkan berbagai kumpulan data. Stabilisasi varians juga dikenal sebagai transformasi stabilisasi varians adalah manfaat tambahan dari transformasi yang membuat dispersi atau variasi dari satu variabel kurang bergantung pada nilai variabel lain.

Tidak diragukan lagi, transformasi penghasil simetri yang paling umum digunakan (walaupun bukan satu-satunya yang mungkin) adalah transformasi pangkat yang didefinisikan oleh dua fungsi yang berkaitan erat.

$$T_1(x) = \begin{cases} x^{\lambda}, & \lambda > 0 \\ \ln(x), & \lambda = 0 \\ -(x^{\lambda}), & \lambda < 0 \end{cases} \tag{2.1}$$

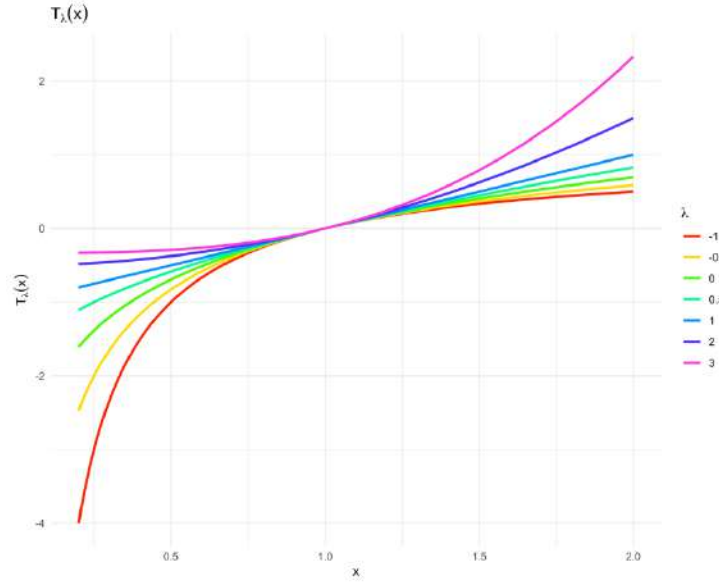
dan

$$T_2(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(x), & \lambda = 0 \end{cases} \tag{2.2}$$

Dalam kasus distribusi variabel data positif yang unimodal, atau single-humped, transformasi ini bermanfaat. Keluarga transformasi didefinisikan oleh masing-masing fungsi ini oleh parameter λ. Nama transformasi pangkat berasal dari fakta bahwa pekerjaan penting dari transformasi ini diselesaikan dengan eksponensial, atau menaikkan nilai data menjadi pangkat λ. Dengan demikian himpunan transformasi dalam Persamaan 1.1 dan 1.2 sebenarnya cukup sebanding dan nilai tertentu λ menghasilkan efek yang sama pada keseluruhan bentuk data dalam kedua kasus tersebut. Transformasi dalam Persamaan 1.1 memiliki bentuk yang sedikit lebih sederhana dan sering digunakan karena lebih mudah. Transformasi dalam Persamaan 1.2, juga dikenal sebagai transformasi Box-Cox adalah versi sederhana dari Persamaan 1.1 yang digeser dan diskalakan, dan terkadang lebih berguna saat membandingkan berbagai transformasi. Juga, Persamaan 1.2 secara matematis lebih baik karena batas transformasi $\lambda \to 0$ sebenarnya adalah fungsi $\ln(x)$.

Untuk transformasi data yang menyertakan beberapa nilai nol atau negatif, rekomendasi dari Box dan Cox (1964) adalah memodifikasi transformasi dengan menambahkan konstanta positif ke setiap nilai data dengan besaran konstanta yang cukup besar untuk semua data yang akan digeser ke setengah

positif dari garis nyata. Pendekatan yang mudah ini seringkali cukup, namun cukup acak. Yeo dan Johnson (2000) telah mengusulkan perluasan terpadu dari transformasi Box-Cox yang mengakomodasi data dimanapun pada garis nyata.



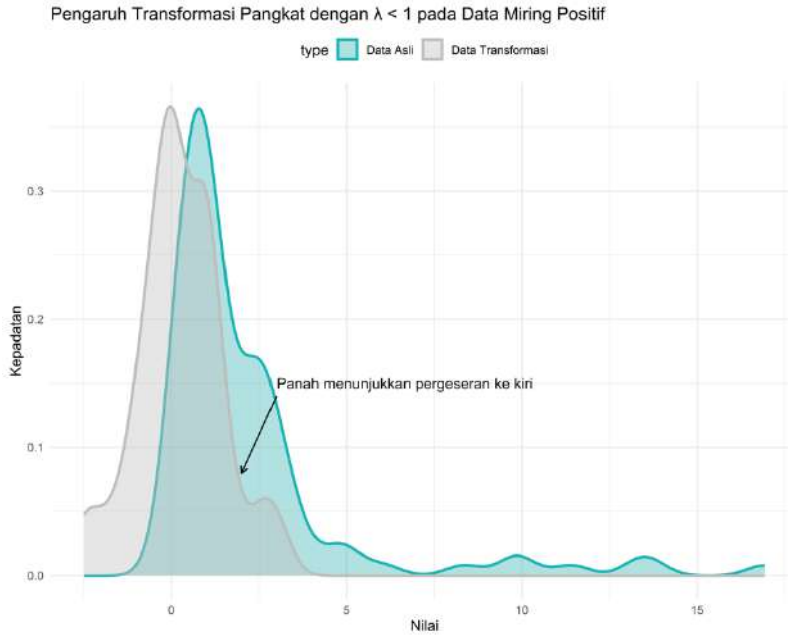
Gambar 2.1 Grafik transformasi pangkat pada Persamaan 1.2 untuk nilai parameter transformasi λ. Untuk $\lambda = 1$ transformasinya linier dan tidak menghasilkan perubahan bentuk data. Untuk $\lambda < 1$ transformasi mengurangi semua nilai data dan nilai yang lebih besar akan lebih terpengaruh. Efek sebaliknya dihasilkan oleh transformasi dengan $\lambda > 1$.

Dalam Persamaan 1.1 dan 1.2, menyesuaikan nilai parameter λ menghasilkan anggota spesifik dari serangkaian transformasi halus yang pada dasarnya terus bervariasi. Transformasi ini terkadang disebut sebagai “tangga pangkat”.

Beberapa dari fungsi transformasi ini diplot pada Gambar 1.1. Kurva pada gambar ini adalah fungsi yang ditentukan oleh Persamaan 1.2, meskipun kurva yang sesuai dari Persamaan 1.1 memiliki bentuk yang sama. **Gambar 2.1** memperjelas bahwa penggunaan transformasi logaritmik untuk $\lambda = 0$ sangat cocok dengan spektrum transformasi pangkat. Gambar ini juga mengilustrasikan sifat lain dari transformasi pangkat, yaitu bahwa semuanya adalah fungsi naik dari variabel awal, x. Sifat ini dicapai dalam Persamaan 1.1 dengan tanda negatif pada transformasi dengan $\lambda < 1$. Untuk transformasi pada Persamaan 1.1 pembalikan tanda ini dicapai dengan membaginya dengan λ. Properti transformasi pangkat yang meningkat secara ketat ini menyiratkan bahwa mereka mempertahankan urutan, sehingga nilai terkecil dalam kumpulan data asli akan sesuai dengan nilai terkecil dalam kumpulan data yang diubah, begitupun dengan nilai terbesar. Faktanya, akan ada korespondensi satu sama lain antara semua statistik urutan dari distribusi asli dan yang diubah. Jadi median, kuartil, dan seterusnya, dari data asli akan menjadi kuantil yang sesuai dari data yang diubah.

Jelas untuk $\lambda = 1$ data pada dasarnya tetap tidak berubah. Untuk $\lambda > 1$ nilai data dinaikkan (kecuali untuk pengurangan $\frac{1}{\lambda}$ dan pembagian dengan λ, jika Persamaan 1.2 digunakan), dengan nilai yang lebih besar dinaikkan lebih banyak daripada yang lebih kecil. Oleh karena itu transformasi pangkat dengan $\lambda > 1$ akan membantu menghasilkan kesimterisan bila diterapkan pada data yang miring negatif.

Kebalikannya berlaku untuk $\lambda < 1$, dimana nilai data yang lebih besar berkurang lebih banyak daripada nilai yang lebih kecil. Oleh karena itu, transformasi pangkat dengan $\lambda < 1$ umumnya diterapkan pada data yang awalnya miring secara positif, untuk menghasilkan data yang hampir simetris.



Gambar 2.2 Pengaruh transformasi pangkat dengan $\lambda < 1$ pada kumpulan data dengan kemiringan positif (kurva tebal). Panah menunjukkan bahwa transformasi memindahkan semua titik ke kiri, dengan nilai yang lebih besar dipindahkan lebih banyak. Distribusi yang dihasilkan (kurva tipis) cukup simetris.

Gambar 2.2 mengilustrasikan mekanisme dari proses ini untuk distribusi yang awalnya miring secara positif (kurva berat). Menerapkan transformasi pangkat dengan $\lambda < 1$

mengurangi semua nilai data, tetapi memengaruhi nilai yang lebih besar dengan lebih kuat. Pilihan yang tepat dari λ seringkali dapat menghasilkan setidaknya perkiraan simetri melalui proses ini. Memilih nilai yang terlalu kecil atau negatif untuk λ akan menghasilkan koreksi berlebih, yang mengakibatkan distribusi yang ditransformasi menjadi miring secara negatif.

Inspeksi awal plot data eksplorasi seperti diagram skematik dapat menunjukkan dengan cepat arah dan perkiraan besarnya kemiringan dalam kumpulan data. Oleh karena itu biasanya jelas apakah transformasi pangkat dengan $\lambda > 1$ atau $\lambda < 1$ sesuai, tetapi nilai spesifik untuk eksponen tidak akan begitu jelas. Sejumlah pendekatan untuk memilih parameter transformasi yang tepat telah disarankan. Yang paling sederhana dari ini adalah statistik d_λ (Hinkley 1977),

$$d_\lambda = \frac{|\text{mean}(\lambda) - \text{median}(\lambda)|}{\text{spread}(\lambda)} \tag{2.3}$$

Di sini, sebaran (*spread*) adalah ukuran dispersi yang resisten, seperti IQR atau MAD. Setiap nilai λ akan menghasilkan rata-rata, median, dan sebaran yang berbeda dalam kumpulan data tertentu, dan ketergantungan pada λ ini ditunjukkan dalam persamaan. Hinkley d_λ digunakan untuk menentukan di antara transformasi pangkat yang dicobakan, dengan menghitung nilainya untuk masing-masing dari sejumlah pilihan yang berbeda untuk λ. Biasanya nilai-nilai percobaan λ diberi jarak dengan interval $1/2$ atau $1/4$. Pemilihan λ

ditentukan dengan d_λ terkecil kemudian diadopsi untuk mengubah data (transformasi data). Salah satu cara yang sangat mudah untuk melakukan perhitungan adalah dengan menggunakan bantuan komputasi.

Dasar dari statistik d_λ adalah untuk data yang terdistribusi secara simetris, rata-rata dan median akan sangat dekat. Oleh karena itu, seiring transformasi pangkat yang lebih kuat secara berturut-turut (nilai λ semakin jauh dari 1) memindahkan data ke arah simetri, pembilang dalam Persamaan 1.3 akan bergerak menuju nol. Karena transformasi menjadi terlalu kuat, pembilang akan mulai meningkat relatif terhadap ukuran sebaran, menghasilkan peningkatan lagi.

Persamaan 1.3 adalah pendekatan sederhana untuk menemukan transformasi pangkat yang menghasilkan simetris atau mendekati simetris dalam data yang ditransformasikan. Pendekatan yang lebih canggih disarankan dalam makalah Box and Cox (1964), yang sangat sesuai ketika data yang diubah harus memiliki distribusi sedekat mungkin dengan Gaussian berbentuk lonceng, misalnya ketika hasil dari beberapa transformasi akan dirangkum secara bersamaan melalui multivariat Gaussian, atau distribusi normal multivariat. Secara khusus, Box dan Cox menyarankan memilih eksponen transformasi pangkat untuk memaksimalkan fungsi log-likelihood untuk distribusi Gaussian

$$L(\lambda) = -\frac{n}{2}\ln[s^2(\lambda)] + (\lambda - 1)\sum_{i=1}^{n}\ln[x_i] \qquad (2.4)$$

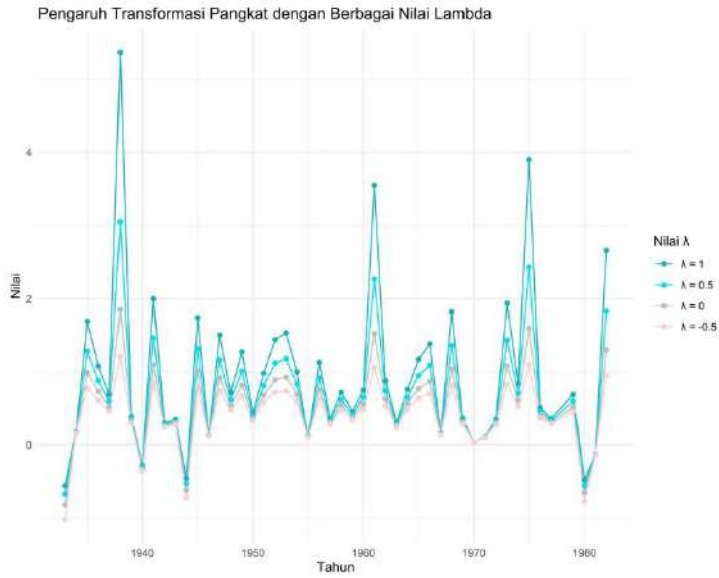
Di sini n adalah ukuran sampel dan s^2 adalah varians sampel (dihitung dengan pembagi n, bukan $n-1$) dari data setelah transformasi dengan eksponen λ. Seperti kasus untuk menggunakan statistik Hinkley (Persamaan 1.3), nilai yang berbeda dari λ dicoba dan yang menghasilkan nilai $L(\lambda)$ terbesar dipilih sebagai yang paling tepat. Ada kemungkinan bahwa dua kriteria akan menghasilkan pilihan yang berbeda karena Persamaan 1.3 hanya membahas simetris data yang diubah, sedangkan Persamaan 1.4 mencoba mengakomodasi semua aspek distribusi Gaussian, termasuk namun tidak terbatas pada simetrinya.

Contoh 2.1 Memilih Transformasi Pangkat yang Tepat

Tabel 1.3 menunjukkan data curah hujan Ithaca Januari 1933–1982 dari Tabel A.2 di Lampiran A, diurutkan dalam urutan menaik dan mengalami transformasi pangkat menggunakan Persamaan 1.2, untuk $\lambda = 1$, $\lambda = 0.5$, $\lambda = 0$, dan $\lambda = -0.5$. Untuk $\lambda = 1$ transformasi hanya berarti mengurangkan 1 dari setiap nilai data. Perhatikan bahwa bahkan untuk eksponen negatif $\lambda = -0.5$ urutan data asli dipertahankan dalam semua transformasi, sehingga mudah untuk menentukan median dan kuartil dari data asli dan data yang diubah.

Tabel 2.1 Curah hujan Ithaca Januari 1933–1982 dari Tabel A.2 ($\lambda = 1$). Data telah disortir dengan transformasi pangkat pada Persamaan 1.2 diterapkan untuk $\lambda = 1$, $\lambda = 0.5$, $\lambda = 0$, dan $\lambda = -0.5$. Plot skematis dari data ini ditunjukkan pada **Gambar 2.3**.

Year	$\lambda = 1$	$\lambda = 0.5$	$\lambda = 0$	$\lambda = -0.5$	Year	$\lambda = 1$	$\lambda = 0.5$	$\lambda = 0$	$\lambda = -0.$
1933	-0.56	-0.67	-0.82	-1.02	1948	0.72	0.62	0.54	0.48
1980	-0.48	-0.56	-0.65	-0.77	1960	0.75	0.65	0.56	0.49
1944	-0.46	-0.53	-0.62	-0.72	1964	0.76	0.65	0.57	0.49
1940	-0.28	-0.30	-0.33	-0.36	1974	0.84	0.71	0.61	0.53
1981	-0.13	-0.13	-0.14	-0.14	1962	0.88	0.74	0.63	0.54
1970	0.03	0.03	0.03	0.03	1951	0.98	0.81	0.68	0.58
1971	0.11	0.11	0.10	0.10	1954	1.00	0.83	0.69	0.59
1955	0.12	0.12	0.11	0.11	1936	1.08	0.88	0.73	0.61
1946	0.13	0.13	0.12	0.12	1956	1.13	0.92	0.76	0.63
1967	0.16	0.15	0.15	0.14	1965	1.17	0.95	0.77	0.64
1934	0.18	0.17	0.17	0.16	1949	1.27	1.01	0.82	0.67
1942	0.30	0.28	0.26	0.25	1966	1.38	1.09	0.87	0.70
1963	0.31	0.29	0.27	0.25	1952	1.44	1.12	0.89	0.72
1943	0.35	0.32	0.30	0.28	1947	1.50	1.16	0.92	0.74
1972	0.35	0.32	0.30	0.28	1953	1.53	1.18	0.93	0.74
1957	0.36	0.33	0.31	0.29	1935	1.69	1.28	0.99	0.78
1969	0.36	0.33	0.31	0.29	1945	1.74	1.31	1.01	0.79
1977	0.36	0.33	0.31	0.29	1939	1.82	1.36	1.04	0.81
1968	0.39	0.36	0.33	0.30	1950	1.82	1.36	1.04	0.81
1973	0.44	0.40	0.36	0.33	1959	1.94	1.43	1.08	0.83
1941	0.46	0.42	0.38	0.34	1976	2.00	1.46	1.10	0.85
1982	0.51	0.46	0.41	0.37	1937	2.66	1.83	1.30	0.95
1961	0.69	0.60	0.52	0.46	1979	3.55	2.27	1.52	1.06
1975	0.69	0.60	0.52	0.46	1958	3.90	2.43	1.59	1.10
1938	0.72	0.62	0.54	0.48	1978	5.37	3.05	1.85	1.21



Gambar 2.3 Pengaruh transformasi pangkat dalam Persamaan 1.2 pada data curah hujan total bulan Januari untuk Ithaca, 1933–1982 (Tabel A.2). Data asli ($\lambda = 1$) sangat miring ke kanan, dengan nilai terbesar berada jauh di luar. Transformasi akar kuadrat ($\lambda = 0.5$) sedikit meningkatkan simetri. Transformasi logaritmik ($\lambda = 0$) menghasilkan distribusi yang cukup simetris. Ketika mengalami transformasi akar kuadrat terbalik yang lebih ekstrem ($\lambda = -0.5$) data mulai menunjukkan kemiringan negatif. Transformasi logaritmik akan dipilih sebagai yang terbaik berdasarkan statistik Hinkley d_λ (Persamaan 1.3), dan log-likelihood Gaussian (Persamaan 1.4).

Gambar 2.3 menunjukkan plot skematik untuk data pada Tabel 1.1. Data yang tidak diubah (plot paling kiri) jelas miring positif, yang biasa dijumpai pada distribusi jumlah curah hujan. Ketiga nilai di luar pagar adalah jumlah yang besar, dengan yang terbesar berada jauh di luar. Tiga plot

skematik lainnya menunjukkan hasil transformasi pangkat yang semakin kuat dengan $\lambda < 1$. Transformasi logaritmik ($\lambda = 0$) keduanya meminimalkan statistik Hinkley d_λ (Persamaan 1.3) dengan IQR sebagai ukuran penyebaran, dan memaksimalkan log-likelihood Gaussian (Persamaan 1.4). Kesimetrisan dekat yang ditunjukkan oleh plot skematik untuk data yang diubah secara logaritmik mendukung kesimpulan bahwa ini adalah yang terbaik di antara kemungkinan yang dipertimbangkan menurut kedua kriteria. Transformasi akar-akar terbalik yang lebih ekstrem ($\lambda = -0.5$) jelas telah dikoreksi berlebihan untuk kemiringan positif, karena tiga jumlah terkecil sekarang berada di luar pagar bawah.

2.1.2 Standarisasi Anomali (Normalisasi)

Meskipun tidak dapat dibandingkan secara menyeluruh, transformasi dapat bermanfaat saat bekerja dengan kumpulan data yang relevan. Situasi di mana data bergantung pada variasi musiman adalah salah satu contoh dari kondisi ini. Misalnya, perbandingan langsung suhu mentah bulanan biasanya akan menentukan siklus musiman. Rekor sejuk Juli akan jauh lebih dingin daripada rekor Januari yang hangat. Dalam keadaan seperti ini, standarisasi anomali dapat sangat membantu dengan penskalaan ulang data.

Standarisasi anomali, z, dihitung hanya dengan mengurangkan rata-rata sampel pada data mentah x, dan membaginya dengan standar deviasi sampel yang sesuai:

$$z = \frac{x - \bar{x}}{s_x} = \frac{x'}{s_x} \tag{2.5}$$

Dalam ilmu atmosfer, anomali x' dipahami sebagai pengurangan dari nilai data rata-rata yang relevan, seperti pada pembilang Persamaan 1.5. Istilah anomali tidak berkonotasi dengan nilai data atau peristiwa yang tidak normal atau bahkan tidak biasa. Standarisasi anomali pada Persamaan 1.5 dihasilkan dengan membagi anomali pada pembilang dengan standar deviasi yang sesuai. Transformasi ini terkadang juga disebut sebagai normalisasi. Dimungkinkan juga untuk membangun Standarisasi anomali menggunakan ukuran lokasi dan penyebaran yang resisten, misalnya, mengurangkan median dan membaginya dengan IQR, namun ini jarang dilakukan. Penggunaan standarisasi anomali dimotivasi oleh ide-ide yang diturunkan dari distribusi Gaussian berbentuk lonceng. Namun, tidak perlu berasumsi bahwa kumpulan data mengikuti distribusi tertentu untuk mengekspresikannya kembali dalam bentuk standarisasi anomali dan mengubah data non-Gaussian menurut Persamaan 1.5 tidak akan membuatnya menjadi Gaussian kembali.

Ide di balik Standarisasi anomali adalah mencoba menghilangkan pengaruh lokasi dan penyebaran dari sekumpulan data. Unit fisik dari data asli dibatalkan, sehingga standarisasi anomali selalu merupakan kuantitas tanpa dimensi. Pengurangan rata-rata menghasilkan serangkaian

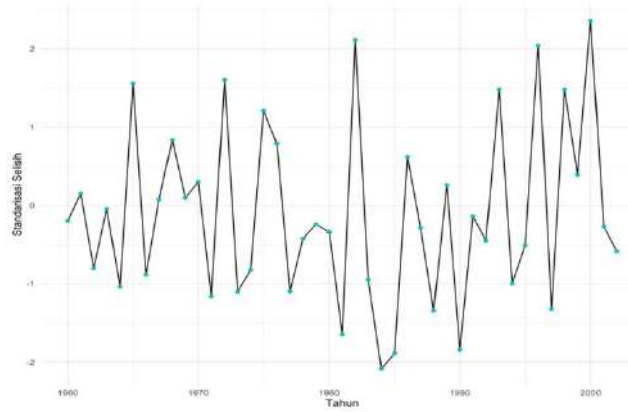
anomali, x', yang terletak di suatu tempat mendekati nol. Pembagian dengan standar deviasi menempatkan ekskursi dari rata-rata dalam kumpulan data yang berbeda pada pijakan yang sama. Secara kolektif, kumpulan data yang telah diubah menjadi sekumpulan standarisasi anomali akan menunjukkan rata-rata nol dan standar deviasi 1.

Misalnya, sering ditemukan bahwa suhu musim panas kurang bervariasi dibandingkan suhu musim dingin. Mungkin ditemukan bahwa standar deviasi untuk suhu rata-rata bulan Januari di beberapa lokasi adalah sekitar 3°C, tetapi standar deviasi untuk suhu rata-rata bulan Juli di lokasi yang sama mendekati 1°C. Suhu rata-rata bulan Juli 3°C lebih dingin daripada rata-rata jangka panjang untuk bulan Juli akan sangat tidak biasa, sesuai dengan standarisasi anomali standar -3. Suhu rata-rata Januari 3°C lebih hangat daripada suhu rata-rata Januari jangka panjang di lokasi yang sama akan menjadi kejadian yang cukup biasa, sesuai dengan standarisasi anomali $+1$. Cara lain untuk melihat standarisasi anomali adalah sebagai ukuran jarak dalam satuan standar deviasi antara nilai data dan rata-ratanya.

Contoh 2.2 Mengekspresikan Data Iklim dalam Standarisasi Anomali

Gambar 2.3 mengilustrasikan penggunaan standarisasi anomali dalam konteks operasional. Titik yang diplot adalah nilai indeks Indeks Osilasi Selatan yang merupakan indeks fenomena El-Niño-Osilasi Selatan (ENSO) yang digunakan

oleh Pusat Prediksi Iklim dari Pusat Prediksi Lingkungan Nasional AS (Ropelewski dan Jones, 1987). Nilai indeks ini dalam gambar diperoleh dari perbedaan bulan demi bulan dalam standarisasi anomali tekanan permukaan laut di dua lokasi tropis: Tahiti, di tengah Samudra Pasifik; dan Darwin, di Australia utara. Dalam hal Persamaan 1.5, langkah pertama untuk menghasilkan **Gambar 2.3** adalah menghitung perbedaan $\Delta z = z_{Tahiti} - z_{Darwin}$ untuk setiap bulan selama tahun-tahun yang diplot. Standarisasi anomali z_{Tahiti} untuk Januari 1960, misalnya, dihitung dengan mengurangkan tekanan rata-rata untuk semua Januari di Tahiti dari tekanan bulanan yang diamati untuk Januari 1960. Perbedaan ini kemudian dibagi dengan standar deviasi yang mencirikan variasi tahun-ke-tahun dari Tekanan atmosfer Januari di Tahiti.



Gambar 2.4 Standarisasi selisih antar standarisasi anomali tekanan permukaan laut bulanan di Tahiti dan Darwin (Indeks Osilasi Selatan), 1960–2002. Nilai bulanan individu telah dihaluskan dalam waktu.

Sebenarnya, kurva pada **Gambar 2.4** didasarkan pada nilai bulanan yang merupakan standarisasi anomali dari perbedaan standarisasi anomali Δz, sehingga Persamaan 1.5 telah diterapkan dua kali pada data asli. Yang pertama dari dua standarisasi dilakukan untuk meminimalkan pengaruh perubahan musiman pada tekanan bulanan rata-rata dan variabilitas tekanan bulanan dari tahun ke tahun. Standardisasi kedua, menghitung standarisasi anomali dari perbedaan Δz, memastikan bahwa indeks yang dihasilkan akan memiliki satuan standar deviasi.

Secara fisik, selama peristiwa El Niño, pusat aktivitas curah hujan Pasifik tropis bergeser ke arah timur dari Pasifik barat (dekat Darwin) ke Pasifik tengah (dekat Tahiti). Pergeseran ini dikaitkan dengan tekanan permukaan yang lebih tinggi dari rata-rata di Darwin dan tekanan permukaan yang lebih rendah dari rata-rata di Tahiti, yang bersama-sama menghasilkan nilai negatif untuk indeks yang diplot pada **Gambar 2.4**. Peristiwa El-Niño yang sangat kuat pada tahun 1982–1983 sangat menonjol dalam gambar ini.

2.2 Teknik Eksplorasi Data Berpasangan

Bab ini sejauh ini menyajikan metode terutama untuk manipulasi dan investigasi kumpulan data tunggal. Plot skematik berdampingan yang ditunjukkan pada **Gambar 2.1** adalah salah satu contoh perbandingan yang telah dibuat. Beberapa distribusi data dari Lampiran A diplot di sini, tetapi tidak menampilkan elemen struktur data yang mungkin

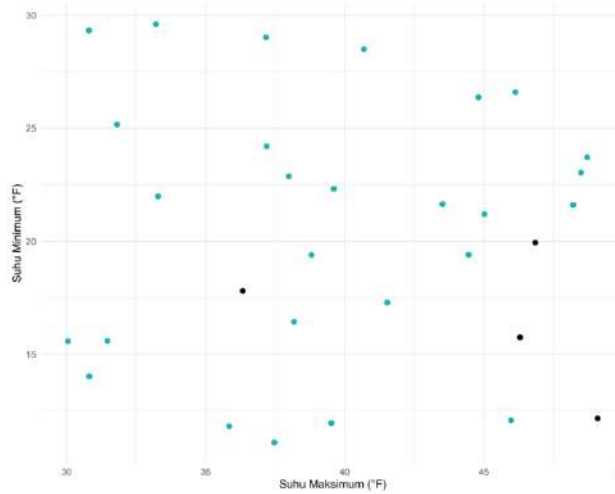
signifikan. Sebelum pembuatan plot skematik, data dari setiap kumpulan diberi peringkat secara terpisah, hubungan antara variabel yang diamati pada hari tertentu disamarkan. Namun, observasi dari satu kelompok terkait dengan yang lain pada tanggal yang sama. Di sini, pengamatan digabungkan. Seringkali, informasi penting dapat diperoleh dengan menjelaskan bagaimana kumpulan pasangan data berinteraksi satu sama lain..

2.2.1 Scatterplot

Format yang hampir universal untuk menampilkan data berpasangan secara grafis adalah scatterplo. Secara geometris, scatterplot hanyalah kumpulan titik di bidang yang dua koordinat Kartesiusnya adalah nilai dari setiap anggota pasangan data. Scatterplot memungkinkan pemeriksaan yang mudah dari fitur-fitur seperti dalam data sebagai tren, kelengkungan dalam hubungan, pengelompokan satu atau kedua variabel, perubahan penyebaran satu variabel sebagai fungsi dari yang lain, dan outlier.

Gambar 2.5 adalah scatterplot dari suhu maksimum dan minimum untuk Ithaca selama Januari 1987. Terlihat bahwa maxima yang sangat dingin diasosiasikan dengan minima yang sangat dingin, dan ada kecenderungan maxima yang lebih hangat diasosiasikan dengan minima yang lebih hangat. Scatterplot ini juga menunjukkan bahwa kisaran tengah suhu maksimum tidak terkait kuat dengan suhu minimum, karena

maksimum mendekati 30°F terjadi dengan minimum dimana saja dalam kisaran −5°F hingga 20°F, atau lebih hangat.



Gambar 2.5 Scatterplot untuk suhu maksimum dan minimum harian selama Januari 1987 di Ithaca, New York. Lingkaran tertutup mewakili hari dengan curah hujan setidaknya 0.01 inci (setara cairan).

Dalam **Gambar 2.5** terlihat juga penggunaan lebih dari satu jenis simbol plotting. Di sini titik-titik yang mewakili hari-hari dimana setidaknya 0.01 inci (setara cairan) dari presipitasi dicatat diplot menggunakan lingkaran yang diisi. Seperti yang terbukti dalam Contoh 2.1 tentang peluang bersyarat, hari hujan cenderung dikaitkan dengan suhu minimum yang lebih hangat. Scatterplot menunjukkan bahwa suhu maksimum juga cenderung lebih hangat, tetapi efeknya tidak terlalu terasa.

2.2.2 Korelasi Pearson

Ukuran hubungan antara dua variabel yang disingkat dan bernilai tunggal, mis. x dan y, wajib. Dalam situasi seperti itu, analis data hampir secara otomatis (dan terkadang tidak kritis) menghitung koefisien korelasi. Umumnya, istilah koefisien korelasi digunakan untuk menunjukkan "koefisien momen produk Pearson dari korelasi linier" antara dua variabel x dan y

Salah satu cara untuk melihat korelasi Pearson adalah rasio sampel kovarians dari dua variabel terhadap hasil dari dua standar deviasi.

$$r_{xy} = \frac{\mathrm{Cov}(x,y)}{s_x s_y}$$

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{1}{2}}\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}}$$

$$= \frac{\sum_{i=1}^{n}(x_i' y_i')}{[\sum_{i=1}^{n}(x_i')^2]^{1/2}[\sum_{i=1}^{n}(y_i')^2]^{1/2}} \qquad (2.6)$$

di mana apostrof menunjukkan anomali atau pengurangan sarana, seperti sebelumnya. Perhatikan bahwa varians sampel adalah kasus khusus dari kovarians (pembilang dalam

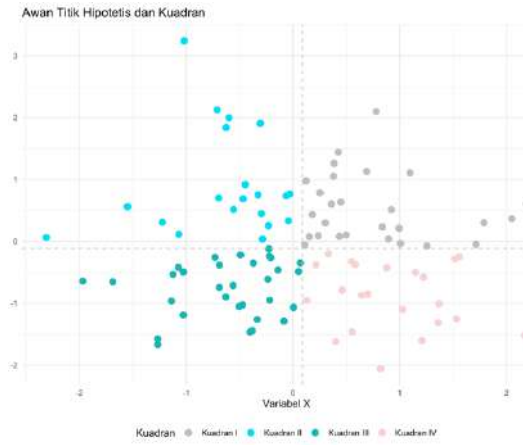
Persamaan 1.7) dengan $x = y$. Salah satu penerapan kovarians adalah dalam matematika untuk menggambarkan turbulensi, di mana produk rata-rata dari, misalnya, anomali kecepatan horizontal u' dan v' disebut kovarians eddy dan digunakan dalam konteks rata-rata Reynolds (mis. Stull, 1988).

Koefisien korelasi momen-produk Pearson tidak kuat atau tahan. Ini tidak kuat tetapi hubungan non-linear antara dua variabel x dan y mungkin tidak dikenali. Itu tidak tahan, karena bisa sangat sensitif terhadap satu atau beberapa pasangan titik luar. Namun demikian, ini digunakan secara luas, baik karena bentuknya sangat cocok untuk manipulasi matematis dan karena terkait erat dengan analisis regresi (lihat Bagian 6.2) dan bivariat (Persamaan 2.33) dan multivariat (lihat Bab 10) distribusi Gaussian.

Korelasi Pearson memiliki dua sifat penting. Pertama, dibatasi oleh -1 dan 1 yaitu $1 \leq r_{xy} \leq 1$. Ketika r_{xy} adalah -1 artinya ada asosiasi linier negatif sempurna antara x dan y. Artinya, diagram sebar y versus x terdiri dari semua titik yang terletak di sepanjang garis, dan garis tersebut memiliki kemiringan negatif. Demikian juga, pada $r_{xy} = 1$ terdapat asosiasi linier positif sempurna. (Perhatikan, bagaimanapun, bahwa $|r_{xy}| = 1$ tidak mengatakan apa-apa tentang kemiringan hubungan linier sempurna antara x dan y, kecuali bahwa itu bukan nol.) Sifat penting kedua adalah bahwa kuadrat dari korelasi Pearson, r_{xy}^2 , adalah pecahan variabilitas menunjukkan salah satu dari dua variabel yang dianggap atau

dijelaskan secara linear oleh yang lain. r_{xy}^2 kadang-kadang dikatakan sebagai fraksi varians dari variabel yang "dijelaskan" oleh yang lain, tetapi interpretasi ini paling tidak tepat dan terkadang menyesatkan. Koefisien korelasi tidak memberikan penjelasan tentang hubungan antara variabel x dan y, setidaknya tidak dalam arti fisik atau kausal. Mungkin saja x secara fisik menyebabkan y atau sebaliknya, tetapi seringkali keduanya dihasilkan secara fisik dari beberapa besaran atau proses lain atau banyak lainnya.

Inti dari koefisien korelasi Pearson adalah kovarians antara x dan y dalam pembilang Persamaan 1.7. Penyebutnya praktis hanya konstanta penskalaan dan selalu positif. Jadi, korelasi Pearson pada dasarnya adalah kovarian tanpa dimensi. Pertimbangkan awan hipotetis dari titik data (x, y) pada **Gambar 2.6**, yang langsung terlihat sebagai korelasi positif. Dua garis tegak lurus melewati dua sarana pemindaian menentukan empat kuadran, biasanya dilambangkan dengan angka Romawi. Untuk setiap n titik, ukuran dalam penjumlahan dihitung dalam pembilang persamaan 1.7. Untuk titik-titik di kuadran I, baik x maupun y nilai lebih besar dari rata-ratanya $(x' > 0 \text{ dan } y' > 0)$, sehingga perkalian kedua faktor tersebut bernilai positif. Oleh karena itu, saya menyumbangkan poin di kuadran secara positif ke persyaratan.



Gambar 2.6 Awan titik hipotetis dalam dua dimensi yang menggambarkan mekanisme koefisien korelasi Pearson (Persamaan 1.7). Dua sampel berarti membagi bidang menjadi empat kuadran, diberi nomor I–IV.

Jumlahkan dalam pembilang persamaan 1.7. Demikian pula, untuk titik-titik di kuadran III, baik x dan y lebih kecil dari rata-ratanya masing-masing ($x' < 0$ dan $y' < 0$), dan sekali lagi produk anomalinya adalah positif. Dengan demikian, poin di kuadran III juga berkontribusi positif terhadap total di pembilang. Untuk poin di kuadran II dan IV, salah satu dari dua variable x dan y di atas rata-rata dan yang lainnya di bawah. Oleh karena itu, hasil kali dalam pembilang Persamaan 1.7 akan negatif untuk titik-titik di kuadran II dan IV, dan titik-titik ini akan memberikan suku negatif pada penjumlahan.

Pada **Gambar 2.6**, sebagian besar titik berada di kuadran I atau III, dan karena itu sebagian besar suku pembilang Persamaan 1.7 adalah positif. Hanya dua titik di kuadran II dan IV yang memberikan suku negatif, dan ini bernilai absolut kecil karena nilai x dan y relatif dekat dengan rata-rata masing-masing. Hasilnya adalah jumlah positif pada pembilangnya dan karenanya merupakan kovarians positif. Dua simpangan baku dalam penyebut Persamaan 1.7 harus selalu positif, memberikan koefisien korelasi positif secara keseluruhan untuk titik-titik pada **Gambar 2.6**. Jika sebagian besar titik berada di kuadran II dan IV, awan titik akan miring ke bawah daripada ke atas, dan koefisien korelasinya akan negatif. Jika awan titik terdistribusi kurang lebih merata di keempat kuadran, koefisien korelasi akan mendekati nol karena suku positif dan negatif dalam penjumlahan pembilang Persamaan 1.5 akan cenderung meniadakan.

Cara lain untuk melihat koefisien korelasi Pearson adalah dengan menggeser konstanta penskalaan dalam penyebut (deviasi standar) dalam jumlah pembilangnya. Operasi ini menghasilkan

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y}\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i} \qquad (2.7)$$

jadi cara lain untuk melihat korelasi Pearson adalah produk rata-rata (mendekati) dari variabel setelah konversi ke anomali standar.

Dari sudut pandang ekonomi komputasi, rumus yang disajikan sejauh ini untuk korelasi Pearson tidak praktis. Ini berlaku terlepas dari apakah perhitungan dilakukan dengan tangan atau dengan program komputer atau tidak. Secara khusus, mereka semua membutuhkan dua lintasan melalui kumpulan data sebelum hasilnya diperoleh: yang pertama untuk menghitung rata-rata sampel, dan yang kedua untuk mengakumulasi suku-suku yang melibatkan penyimpangan nilai data dari milik mereka.

Sampel berarti (anomali). Iterasi melalui kumpulan data dua kali membutuhkan upaya dua kali lipat dan menimbulkan dua kali peluang kesalahan input saat menggunakan kalkulator genggam, dan dapat menambah waktu komputer secara signifikan saat bekerja dengan kumpulan data besar. Oleh karena itu, seringkali berguna untuk mengetahui bentuk perhitungan korelasi Pearson, yang memungkinkannya dihitung hanya dengan satu kali melewati kumpulan data.

Bentuk perhitungan muncul dari manipulasi aljabar sederhana dari penjumlahan dalam koefisien korelasi. Perhatikan pembilang pada Persamaan 1.7. Melakukan perkalian yang ditentukan memberi

$$\sum_{i=1}^{n}[(x_i-\bar{x})(y_i-\bar{y})]=\sum_{i=1}^{n}[x_iy_i-x_i\bar{y}-y_i\bar{x}+\bar{x}\bar{y}]$$

$$=\sum_{i=1}^{n}(x_iy_i)-\bar{y}\sum_{i=1}^{n}x_i-\bar{x}\sum_{i=1}^{n}y_i+\bar{x}\bar{y}\sum_{i=1}^{n}(1)$$

$$=\sum_{i=1}^{n}(x_iy_i)-n\bar{x}\bar{y}-n\bar{x}\bar{y}+n\bar{x}\bar{y}$$

$$=\sum_{i=1}^{n}(x_iy_i)-\frac{1}{n}\left[\sum_{i=1}^{n}x_i\right]\left[\sum_{i=1}^{n}y_i\right] \qquad (2.8)$$

Baris kedua dalam Persamaan 1.9 dicapai dengan mengakui bahwa rata-rata sampel adalah konstan setelah nilai data individual ditentukan, dan karenanya dapat digeser (difaktorkan) di luar penjumlahan. Pada suku terakhir dari baris ini, tidak ada yang tersisa dalam penjumlahan selain angka 1, dan jumlah n-nya adalah n. Langkah ketiga mengakui bahwa ukuran sampel dikalikan dengan rata-rata sampel memberikan jumlah nilai data , diambil langsung dari Definisi mean sampel berikut (Persamaan 1.2). Langkah keempat hanya menggantikan definisi rata-rata sampel lagi untuk menekankan bahwa semua kuantitas yang diperlukan untuk menghitung pembilang korelasi Pearson dapat diketahui setelah melewati satu data. Ini adalah jumlah dari x, jumlah dari y dan jumlah produk mereka.

Dari bentuk penjumlahan yang serupa dalam penyebut korelasi Pearson, harus jelas bahwa rumus analog dapat diturunkan untuk ini atau, dengan kata lain, untuk standar deviasi sampel. Mekanisme penurunannya persis seperti yang diikuti pada Persamaan 1.9, dengan hasil:

$$s_x = \left[\frac{\sum x_i^2 - n\bar{x}^2}{n-1}\right]^{1/2} = \left[\frac{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}{n-1}\right]^{1/2} \qquad (2.9)$$

Hasil yang serupa tentu saja diperoleh untuk y. Secara matematis, Persamaan 1.10 sama persis dengan rumus standar deviasi sampel pada Persamaan 1.6. Dengan demikian, Persamaan 1.9 dan 1.10 dapat disubstitusi ke dalam bentuk korelasi Pearson yang diberikan pada Persamaan 1.7 atau 1.8 untuk memperoleh bentuk perhitungan koefisien korelasi.

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\right]^{1/2}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y\ \ )^2\right]^{1/2}} \qquad (2.10)$$

Bentuk perhitungan untuk koefisien skewness sampel (Persamaan 1.9) menghasilkan cara yang analog.

$$\gamma = \frac{\frac{1}{n-1}\left[\sum x_i^3 - \frac{3}{n}(\sum x_i)(\sum x_i^2) + \frac{2}{n}(\sum x_i)^3\right]}{s^3} \qquad (2.11)$$

Penting untuk menyebutkan peringatan mengenai bentuk perhitungan yang baru saja diturunkan. Penggunaannya

datang dengan potensi masalah yang berasal dari fakta bahwa mereka sangat sensitif terhadap kesalahan pembulatan. Masalah muncul karena masing-masing rumus ini melibatkan selisih antara dua angka yang mungkin besarnya sebanding. Sebagai ilustrasi, asumsikan bahwa dua suku pada baris terakhir Persamaan 1.9 masing-masing telah disimpan menjadi lima angka penting. Jika tiga digit pertama adalah sama, selisihnya hanya diketahui hingga dua digit signifikan, bukan lima. Solusi untuk masalah potensial ini adalah mempertahankan sebanyak mungkin (sebaiknya semua) digit signifikan dalam setiap perhitungan, misalnya dengan menggunakan representasi presisi ganda saat memprogram perhitungan titik-mengambang pada komputer.

Contoh 3.6 Beberapa keterbatasan korelasi linier

Pertimbangkan dua set data buatan pada Tabel 1.4. Nilai data sedikit dan cukup kecil sehingga bentuk perhitungan korelasi Pearson dapat digunakan tanpa membuang angka signifikan. Untuk himpunan I korelasi Pearson $r_{xy} = 0.88$ dan untuk himpunan II korelasi Pearson $r_{xy} = 0{,}61$. Dengan demikian, hubungan linier yang cukup kuat tampak diindikasikan untuk kedua set data berpasangan.

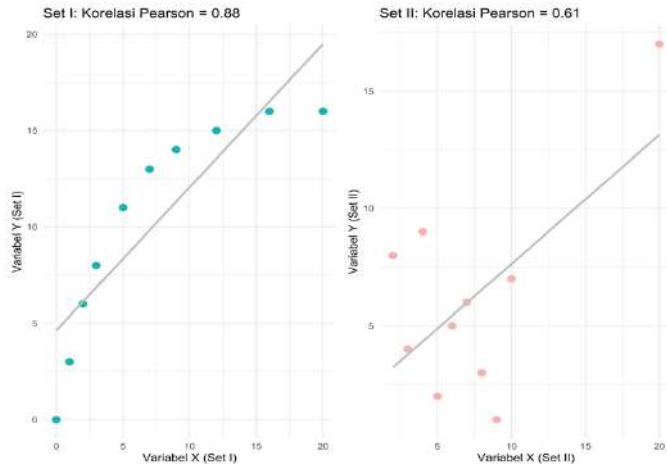
Korelasi Pearson tidak kuat atau persisten, dan dua kumpulan data kecil ini dibuat untuk mengilustrasikan kekurangan ini. **Gambar 2.6** menunjukkan scatterplot dari kedua dataset, dengan set I pada panel (a) dan set II pada panel (b). Untuk Set I, hubungan antara x dan y sebenarnya lebih kuat dari

yang ditunjukkan oleh korelasi linier sebesar 0,88. Semua titik data jatuh hampir pada kurva yang halus, tetapi karena kurva itu bukan garis lurus, koefisien Pearson meremehkan kekuatan hubungan tersebut. Itu tidak kuat untuk penyimpangan dari linearitas dalam suatu hubungan.

Gambar 1.2b menunjukkan bahwa koefisien korelasi Pearson tidak resisten terhadap data eksternal. Terlepas dari titik tepi tunggal, data di Set II menunjukkan sangat sedikit.

Tabel 2.2 Set data berpasangan artifisial untuk contoh korelasi.

Set I		Set II	
x	y	x	y
0	0	2	8
1	3	3	4
2	6	4	9
3	8	5	2
5	11	6	5
7	13	7	6
9	14	8	3
12	15	9	1
16	16	10	7
20	16	20	17



Gambar 2.7 Scatterplot dari dua set data berpasangan buatan pada Tabel 1.4. Korelasi Pearson untuk data pada panel (a) (Set I pada Tabel 1.4) hanya sebesar 0,88 mewakili kekuatan hubungan yang kurang terwakili, menunjukkan bahwa ukuran korelasi ini tidak kuat. Korelasi Pearson untuk data di panel (b) (Teorema II) adalah 0,61, yang mencerminkan pengaruh luar biasa dari titik eksterior tunggal dan menggambarkan tidak adanya perlawanan.

Struktur. Jika ada, sembilan poin yang tersisa ini berkorelasi negatif secara lemah. Namun, nilai $x = 20$ dan $y = 17$ sangat jauh dari sampel masing-masing berarti bahwa produk dari dua perbedaan positif besar yang dihasilkan pada pembilang Persamaan 1.7 atau Persamaan 1.8 mendominasi jumlah keseluruhan, secara salah menunjukkan hubungan positif yang cukup kuat antara sepuluh pasangan data secara total.

Contoh 3.7 Perbandingan korelasi Spearman dan Kendall untuk data pada Tabel 1.4

Di Set I Tabel 1.4, ada hubungan monoton antara x dan y, sehingga masing-masing dari dua tumpukan data sudah dalam urutan menaik. Oleh karena itu, kedua anggota dari masing-masing n pasangan memiliki peringkat yang sama di dalam tumpukan mereka sendiri dan selisih D_i semuanya nol. Faktanya, dua nilai y terbesar adalah sama dan masing-masing akan diberi peringkat 9,5. Kecuali untuk seri ini, jumlah pembilang suku kedua pada Persamaan 1.3 adalah nol dan korelasi pangkat Spearman pada dasarnya adalah 1. Hasil ini mencerminkan kekuatan hubungan antara x dan y lebih baik daripada korelasi Pearson sebesar 0,88. Dengan demikian, koefisien korelasi Pearson mencerminkan kekuatan hubungan linier, tetapi korelasi peringkat Spearman mencerminkan kekuatan hubungan monoton.

Karena data dalam Himpunan I memiliki hubungan monoton positif yang pada dasarnya sempurna, semua $10(10\ 1)/2 = 45$ kemungkinan kecocokan antara pasangan data menghasilkan hubungan yang selaras. Untuk kumpulan data dengan hubungan monoton yang sangat negatif (salah satu variabel benar-benar berkurang sebagai fungsi dari yang lain), semua perbandingan antara pasangan data mengungkapkan hubungan sumbang. Dengan pengecualian dasi, semua perbandingan untuk Himpunan I adalah hubungan yang sesuai. $N_C = 45$, sehingga Persamaan 1.4 akan menghasilkan $\tau(45 - 0)/45 = 1$.

Untuk data pada Set II, nilai-x disajikan dalam urutan menaik, tetapi nilai-y yang dipasangkan dengannya kacau. Perbedaan peringkat untuk set pertama adalah $D_1 = 1 - 8 = -7$. Hanya ada tiga pasangan data di set II yang cocok dengan peringkat (kelima, keenam, dan outlier dari pasangan kesepuluh). Tujuh pasangan yang tersisa akan menyumbangkan suku-suku bukan nol pada penjumlahan dalam Persamaan 1.3, memberikan $r_{rank} = 0.018$ untuk Teorema II. Hasil ini mencerminkan hubungan keseluruhan yang sangat lemah antara x dan y pada Set II jauh lebih baik daripada korelasi Pearson sebesar 0,61.

Menghitung τ Kendall untuk Set II menjadi lebih mudah dengan mengurutkannya dengan meningkatkan nilai variabel x. Dengan pengaturan ini, jumlah kombinasi yang cocok dapat ditentukan dengan menghitung jumlah variabel y berikutnya yang lebih besar dari masing-masing daftar pertama sampai $(n-1)^{st}$ dalam tabel. Secara khusus, ada dua variabel y lebih besar dari 8 dari (2, 8) di antara sembilan nilai di bawah ini, lima variabel y. lebih besar dari 4 dalam (3, 4) di antara delapan nilai di bawah, variabel y lebih besar dari 9 dalam (4, 9) di antara tujuh nilai di bawah, ..., dan variabel y lebih besar dari 7 dalam (10, 7) dalam nilai individu di bawah ini. Semuanya adalah 2 + 5 + 1 + 5 + 3 + 2 + 2 + 2 + 1 = 23 kombinasi konkordan dan 45 − 23 = 22 kombinasi sumbang, memberikan τ = (23 −22)/45 = 0 022.

2.2.3 Korelasi Spearman Rank dan Kendall τ

Tersedia alternatif yang kuat dan tangguh untuk koefisien korelasi momen produk Pearson. Yang pertama dikenal sebagai koefisien korelasi peringkat Spearman. Korelasi Spearman hanyalah koefisien korelasi Pearson yang dihitung menggunakan peringkat data. Secara konseptual, Persamaan 1.7 atau Persamaan 1.8 berlaku, tetapi untuk peringkat data daripada nilai data itu sendiri. Misalnya, pertimbangkan pasangan data pertama (2, 8) di Set II dari Tabel 1.4. Di sini $x = 2$ adalah yang terkecil dari 10 nilai x dan karenanya memiliki peringkat 1. Terkecil kedelapan dari 10 nilai, $y = 8$ memiliki peringkat 8. Dengan demikian, pasangan data pertama ini akan diubah menjadi (1,8) sebelum menghitung korelasi. Demikian pula, nilai x dan y pada pasangan terluar (20,17) adalah yang terbesar dari tumpukan masing-masing 10 dan akan diubah menjadi (10,10) sebelum menghitung koefisien korelasi Spearman.

Dalam prakteknya tidak perlu menggunakan Persamaan 1.7, 1.8 atau 1.11 untuk menghitung Korelasi peringkat Spearman. Sebaliknya, perhitungannya disederhanakan karena kita tahu sebelumnya seperti apa nilai yang diubah. Karena datanya adalah peringkat, mereka hanya terdiri dari semua bilangan bulat dari 1 hingga ukuran sampel n. Misalnya, rata-rata peringkat dari masing-masing empat seri data pada Tabel 1.4 adalah $(1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10)/10 = 5.5$. Demikian pula, simpangan baku (Persamaan 1.10) dari sepuluh bilangan bulat positif pertama ini adalah sekitar

3,028. Secara lebih umum, rata-rata bilangan bulat dari 1 sampai n adalah $(n+1)/2$ dan variansnya adalah $n(n^2-1)/[12(n-1)]$. Dengan menggunakan informasi ini, perhitungan korelasi peringkat Spearman dapat disederhanakan.

$$r_{rank} = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \qquad (2.12)$$

di mana Di adalah perbedaan peringkat antara pasangan ke-i dari nilai data. Dalam kasus ikatan di mana nilai data tertentu muncul lebih dari sekali, semua nilai yang sama diberi peringkat rata-rata sebelum menghitung D_i.

Kendall's τ adalah alternatif kuat dan tahan lama kedua dari Pearson tradisional Korelasi. Kendall's τ dihitung dengan mempertimbangkan hubungan antara semua pasangan data yang mungkin cocok (x_i, y_i) yang terdapat $n(n-1)/2$ dalam sampel berukuran n. Setiap kesepakatan seperti itu di mana kedua anggota dari satu pasangan lebih besar dari rekan mereka di pasangan lainnya dikatakan sesuai. Misalnya, pasangan (3, 8) dan (7, 83) adalah konkordan karena kedua angka pada pasangan terakhir lebih besar daripada pasangannya pada pasangan sebelumnya. Pertandingan di mana setiap pasangan memiliki salah satu nilai yang lebih besar, misalnya (3, 83) dan (7, 8), dikatakan sumbang. Kendall's τ dihitung dengan mengurangkan jumlah pasangan sumbang, ND, dari jumlah pasangan konkordan, NC, dan

membaginya dengan jumlah kemungkinan kecocokan di antara n pengamatan.

$$\tau = \frac{N_c - N_D}{n(n-1)/2} \tag{2.13}$$

Pasangan identik berkontribusi $1/2$ untuk N_C dan N_D.

2.2.4 Fungsi Autokorelasi

Di Bab 2, probabilitas bersyarat untuk dua kejadian berbeda, presipitasi dan tidak presipitasi, digambarkan untuk persistensi meteorologi, atau kecenderungan cuaca yang serupa selama periode berturut-turut. Persistentsi variabel kontinu, seperti suhu, biasanya digambarkan sebagai korelasi serial atau autokorelasi temporal. Autokorelasi temporal menunjukkan hubungan suatu variabel dengan nilai masa depan dan masa lalunya sendiri, seperti yang ditunjukkan oleh awalan "otomatis" dalam autokorelasi. Autokorelasi biasanya dihitung sebagai koefisien korelasi momen produk Pearson, tetapi tidak ada alasan mengapa jenis korelasi lagged lainnya juga tidak dapat dihitung.

Proses penghitungan autokorelasi dapat divisualisasikan dengan membayangkan bahwa dua salinan dari rangkaian pengamatan ditulis, dengan salah satu rangkaian digeser oleh satu satuan waktu. Pergeseran ini ditunjukkan pada Gambar 1.8 menggunakan data suhu maksimum Ithaca Januari 1987 dari Tabel A.1. Seri data ini telah ditulis ulang dengan bagian tengah bulan diwakili oleh elips di baris pertama. Rekor yang

sama disalin lagi di baris kedua, tetapi suatu hari digeser ke kanan. Proses ini menghasilkan 30 pasang suhu di dalam kotak yang tersedia untuk menghitung koefisien korelasi.

Autokorelasi dihitung dengan mengganti pasangan data tertinggal ke dalam rumus korelasi Pearson (Persamaan 1.7). Untuk autokorelasi lag 1 ada n 1 pasangan seperti itu. Satu-satunya kebingungan yang sebenarnya muncul karena sarana untuk kedua seri tersebut umumnya akan sedikit berbeda. Misalnya, pada **Gambar 2.8**, rata-rata dari 30 nilai kotak di baris atas adalah 29.77°F dan rata-rata nilai kotak di baris bawah kisaran 29.73°F. Perbedaan ini muncul karena baris paling atas tidak memuatnya. Suhu untuk 1 Januari, dan baris terbawah tidak termasuk suhu untuk 31 Januari. Rata-rata sampel dari nilai $n-1$ pertama dilambangkan dengan subskrip "−" dan nilai n − 1 terakhir dengan subskrip "+", yang merupakan autokorelasi lag-1.

$$r_1 = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(x_{i+1} - \bar{x}_+)]}{[\sum_{i=1}^{n-1}(x_i - \bar{x}_-)^2 \sum_{i=2}^{n}(x_i - \bar{x}_+)^2\ ]^{1/2}} \qquad (2.14)$$

Misalnya, untuk data temperatur maksimum dari Ithaca pada Januari 1987, $r_1 = 0.52$.

33	32	30	29	25	30	53	•	•	•	17	26	27	30	34
	33	32	30	29	25	30	53	•	•	•	17	26	27	30	34

Gambar 2.8 Konstruksi rangkaian waktu tergeser dari data suhu maksimum Ithaka untuk Januari 1987. Pergeseran data satu hari

menyisakan 30 pasangan data (terlampir dalam kotak) untuk menghitung koefisien autokorelasi lag 1.

Autokorelasi Lag-1 adalah ukuran persistensi yang paling umum dihitung, tetapi terkadang menarik untuk menghitung autokorelasi pada lag yang lebih panjang. Secara konseptual, ini tidak lebih sulit daripada prosedur untuk autokorelasi lag 1, dan secara komputasi satu-satunya perbedaan adalah bahwa kedua deret digeser lebih dari satu satuan waktu. Tentu saja, karena rangkaian waktu semakin bergeser relatif terhadap dirinya sendiri, semakin sedikit data yang tumpang tindih untuk dikerjakan. Persamaan 1.15 dapat digeneralisasikan ke koefisien autokorelasi lag-k dengan menggunakan.

$$r_k = \frac{\sum_{i=1}^{n-k}[(x_i - \bar{x})(x_{i+k} - \bar{x}_+)]}{[\sum_{i=1}^{n-k}(x_i - \bar{x})^2 \sum_{i=k+1}^{n}(x_i - \bar{x}_+)^2]^{1/2}} \qquad (2.15)$$

Di sini, indeks "-" dan "+" masing-masing menunjukkan rata-rata sampel pada nilai data nk pertama dan terakhir. Persamaan 1.16 berlaku untuk 0 k < n 1, meskipun biasanya hanya beberapa nilai k terkecil yang menarik. Dengan penundaan yang lama, begitu banyak data yang hilang sehingga korelasi untuk katakanlah k > n/2 atau k > n/3 jarang dihitung.

Dalam situasi di mana kumpulan data yang panjang tersedia, terkadang dapat diterima untuk menggunakan perkiraan Persamaan 1.16, yang menyederhanakan perhitungan dan memungkinkan satu bentuk perhitungan digunakan.

Khususnya, jika deret data cukup panjang, rata-rata sampel keseluruhan sangat dekat dengan rata-rata himpunan bagian dari nilai $n-k$ pertama dan terakhir. Deviasi standar sampel massal juga akan mendekati dua standar deviasi subset untuk nilai nk pertama dan terakhir. Mengandalkan asumsi ini mengarah pada perkiraan yang sangat umum digunakan.

$$r_k \approx \frac{\sum_{i=1}^{n-k}[(x_i-\bar{x})(x_{i+k}-\bar{x})]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\approx \frac{\sum_{i=1}^{n}(x_i x_{i+k}) - \frac{n-k}{n^2}(\sum_{i=1}^{n} x_i)^2}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} \qquad (2.16)$$

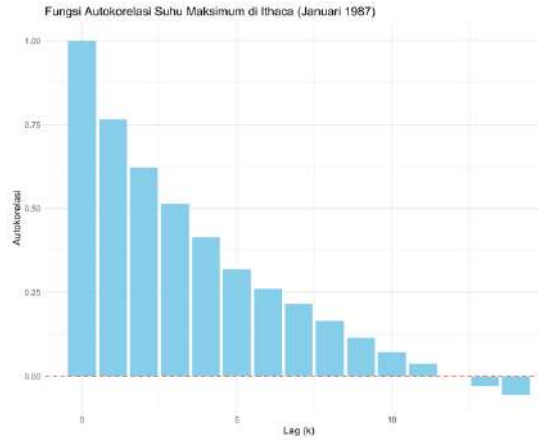
2.2.5 Fungsi Autokorelasi

Fungsi autokorelasi adalah kumpulan autokorelasi yang dihitung untuk beberapa kelambatan. Seringkali, nilai pertama fungsi autokorelasi untuk data suhu maksimum dari Ithaca pada Januari 1987 ditunjukkan dalam **Gambar 2.9**. Karena setiap deret data tidak tergeser memiliki korelasi sempurna dengan dirinya sendiri, fungsi autokorelasi selalu dimulai dengan r_0=1. Fungsi autokorelasi biasanya menunjukkan penurunan yang kurang lebih bertahap menuju nol saat lag k meningkat. Ini menunjukkan hubungan statistik yang umumnya lebih lemah antara titik data yang terpisah lebih jauh dalam waktu. Prakiraan cuaca harus mempertimbangkan pengamatan ini. Akan sangat mudah untuk membuat prediksi yang cukup akurat dalam rentang

waktu ini jika fungsi autokorelasi tidak turun menuju nol setelah beberapa hari: prediksi sederhana dari observasi hari ini (perkiraan persistensi) atau modifikasi dari observasi hari ini akan memberikan hasil yang baik.

Terkadang berguna untuk mengubah skala fungsi autokorelasi dengan mengalikan semuanya autokorelasi melalui varian data. Hasilnya, yang sebanding dengan pembilang Persamaan 1.16 dan 1.17, disebut fungsi autokovarians,

$$\gamma_k = \sigma^2 r_k, k = 0,1,2, \ldots \quad (2.17)$$



Gambar 2.9 Contoh fungsi autokorelasi untuk data suhu maksimum dari Ithaca pada Januari 1987. Korelasinya adalah 1 untuk k = 0 karena data sesaat berkorelasi sempurna dengan diri mereka sendiri. Fungsi autokorelasi turun menjadi hampir nol untuk $k \geq 10$.

Kehadiran autokorelasi dalam data meteorologi dan klimatologi memiliki implikasi penting untuk penerapan beberapa metode statistik standar pada data atmosfer. Secara khusus, penerapan metode klasik yang tidak kritis, yang membutuhkan independensi data dalam sampel, sering menyebabkan hasil yang sangat menyesatkan bila diterapkan pada rangkaian yang sangat persisten. Dalam beberapa kasus, dimungkinkan untuk berhasil memodifikasi teknik ini dengan memperhitungkan ketergantungan waktu menggunakan autokorelasi sampel. Topik ini dibahas dalam Bab 5.

2.3 Teknik Eksplorasi Data Dimensi Tinggi

Metode yang ditunjukkan sejauh ini hanya dapat diterapkan pada himpunan bagian berpasangan dari variabel ketika data yang diperlukan untuk eksplorasi, analisis, atau perbandingan terdiri dari lebih dari dua variabel. Karena kombinasi masalah geometris dan kognitif, menampilkan tiga atau lebih variabel secara bersamaan pada dasarnya sulit. Masalah geometriknya adalah bahwa sebagian besar media tampilan yang tersedia, seperti layar komputer dan kertas, adalah dua dimensi. Oleh karena itu, untuk memplot data berukuran besar secara langsung memerlukan proyeksi geometris ke area di mana informasi proses jelas akan hilang. Masalah kognitif berasal dari fakta bahwa otak kita berevolusi untuk menghadapi kehidupan di dunia tiga dimensi, dan memvisualisasikan empat dimensi atau lebih secara bersamaan itu sulit atau tidak mungkin. Namun demikian, alat grafis pintar untuk EDA

multivariat (tiga atau lebih variabel secara bersamaan) telah dikembangkan. Selain ide yang disajikan di bagian ini, beberapa perangkat EDA grafis multivariat yang dirancang khusus untuk prediksi ansambel ditunjukkan di Bagian 6.6.6, dan pendekatan EDA dimensi tinggi berdasarkan analisis komponen utama dijelaskan di Bagian 11.7.3.

2.3.1 Plot Bintang

Jika jumlah variabel K tidak terlalu besar, setiap himpunan n observasi berdimensi K dapat digambarkan sebagai diagram bintang. Bagan bintang didasarkan pada sumbu koordinat K dengan titik asal yang sama berjarak 360o/K terpisah di bidang. Untuk setiap n observasi, nilai kth dari variabel K sebanding dengan jarak plot radial pada sumbu yang bersesuaian. "Bintang" terdiri dari segmen garis yang menghubungkan titik-titik ini dengan pasangannya pada sumbu radial yang berdekatan.

Gambar 2.10 Bagan bintang selama lima hari terakhir dalam data Januari 1987 pada Tabel A.1, dengan sumbu berlabel hanya untuk bintang 27 Januari. Perkiraan simetri radial dalam bagan ini mencerminkan korelasi antara variabel serupa di dua lokasi, dan perluasan bintang dari waktu ke waktu menunjukkan hari yang lebih hangat dan lebih basah menjelang akhir bulan.

Sebagai contoh, **Gambar 2.10** menunjukkan grafik bintang untuk 5 (dari n = 31) hari terakhir dari data Januari 1987 pada Tabel A.1. Karena ada variabel $K = 6$, enam sumbu dipisahkan oleh sudut 360°/660° dan masing-masing diidentifikasi dengan salah satu variabel seperti yang diberikan pada tablet 27 Januari. Secara umum, skala proporsionalitas pada bagan bintang berbeda untuk variabel yang berbeda dan dirancang sedemikian rupa sehingga nilai terkecil (atau nilai yang mendekati tetapi bawah) sesuai dengan asal, dan nilai terbesar (atau nilai dekat atau di atas) sesuai dengan panjang penuh sumbu. Karena variabel pada **Gambar 2.10** memiliki tipe yang sama, maka skala untuk ketiga tipe variabel tersebut dipilih yang identik agar dapat membandingkannya dengan lebih baik. Misalnya, asal Ithaca dan Canandaigua Sumbu suhu maksimum sesuai dengan 10°F, dan ujung sumbu ini sesuai dengan 40°F. Sumbu presipitasi memiliki nol di titik awal dan 0,15 inci di ujung, sehingga bentuk segitiga ganda untuk 27 Januari dan 28 Januari menunjukkan nol presipitasi di kedua lokasi. untuk hari-hari itu. Simetri dekat bintang menunjukkan korelasi yang kuat untuk pasangan variabel serupa (karena sumbu mereka diplot terpisah 180°), dan kecenderungan bintang menjadi lebih besar dari waktu ke waktu menunjukkan hari yang lebih hangat dan lebih basah di akhir bulan.

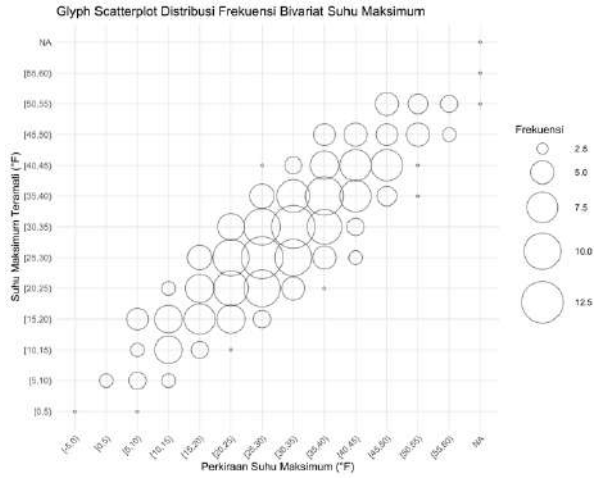
2.3.2 Glyph Scatterplot

Plot glyph scatterplot adalah perluasan dari sebar sebar biasa, di mana titik-titik sederhana yang menemukan titik-titik pada bidang dua dimensi yang ditentukan oleh dua variabel diganti dengan "mesin terbang" atau simbol yang lebih rumit yang mewakili nilai variabel tambahan di dalamnya besaran dan menyandikan ukuran/atau bentuk. Gambar 1.5 adalah sebaran glif primitif di mana glif sirkular terbuka/tertutup menunjukkan variabel presipitasi/tanpa presipitasi biner.

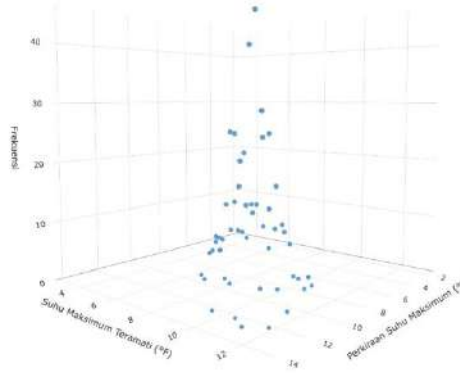
Gambar 1.11 adalah diagram glyph scatterplot sederhana yang menunjukkan tiga variabel yang terkait dengan evaluasi sekelompok kecil prakiraan suhu maksimum musim dingin. Dua sumbu sebar adalah suhu yang diprediksi dan diamati yang dibulatkan menjadi 5°F nampan, dan mesin terbang melingkar digambar sehingga luasnya sebanding dengan jumlah pasangan perkiraan-pengamatan dalam nampan persegi $5°F \times 5°F$ yang diberikan. Memilih area yang sebanding dengan variabel ketiga (di sinilah data di setiap nampan dihitung) lebih disukai daripada radius atau diameter, karena permukaan mesin terbang lebih cocok dengan indra visual ukuran.

Gambar 2.11 pada dasarnya adalah histogram dua dimensi untuk kumpulan data suhu bivariat ini, tetapi lebih efektif daripada generalisasi langsung dari histogram dua dimensi tradisional untuk satu variabel menjadi tiga dimensi. **Gambar 2.12** menunjukkan histogram bivariat seperti itu dalam

tampilan perspektif, yang cukup tidak efektif karena proyeksi tiga dimensi ke halaman dua dimensi telah menimbulkan ambiguitas tentang posisi masing-masing titik. Ini adalah kasusnya, meskipun setiap titik pada **Gambar 2.12** terikat pada posisinya pada tingkat pengamatan yang diperkirakan pada dasar plot yang tampak oleh:

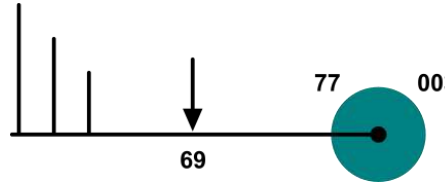


Gambar 2.11 Glyph scatterplot distribusi frekuensi bivariat dari perkiraan dan suhu maksimum musim dingin harian yang diamati untuk Minneapolis, 1980–1981 hingga 1985–1986. Suhu telah dibulatkan ke interval 5°F, dan mesin terbang melingkar telah diskalakan untuk memiliki area yang sebanding dengan hitungan (inset).



Gambar 2.12 Histogram bivariat disajikan dalam tampilan perspektif dari data yang sama disajikan sebagai scatterplot mesin terbang pada **Gambar 2.11**. Meskipun titik data pada bidang prediksi-pengamatan terletak di ekor vertikal dan titik pada diagonal 1:1 selanjutnya ditunjukkan oleh lingkaran terbuka, proyeksi dari tiga ke dua dimensi memperumit interpretasi gambar. Dari Murphy dkk. (1989).

ekor vertikal dan titik yang jatuh tepat pada diagonal ditunjukkan dengan simbol karakter terbuka. **Gambar 2.11** berbicara lebih jelas daripada **Gambar 2.12** tentang data dan, misalnya, segera menunjukkan bahwa ada kecenderungan prediksi yang berlebihan (suhu yang diprediksi, rata-rata, secara sistematis lebih hangat daripada suhu yang diamati). Alternatif yang efektif untuk scatterplot mesin terbang pada **Gambar 2.11** untuk menampilkan bivariat.



Gambar 2.13 Mesin terbang canggih yang dikenal sebagai model stasiun meteorologi yang mewakili tujuh besaran secara bersamaan. Saat diplot pada peta, itu juga menambahkan dua variabel lokasi (lintang dan bujur), meningkatkan dimensi plot menjadi sembilan magnitudo, yang setara dengan sebaran data cuaca mesin terbang.

Distribusi frekuensi dapat berupa plot kontur estimasi densitas kernel bivariat (lihat Bagian 3.3.6) untuk data ini.

Mesin terbang yang lebih canggih daripada lingkaran pada **Gambar 2.11** dapat digunakan untuk menampilkan data dengan lebih dari tiga variabel sekaligus. Misalnya, mesin terbang bintang dapat digunakan sebagai simbol karakter dalam plot pencar mesin terbang, seperti yang dijelaskan di Bagian 3.6.1. Hampir semua bentuk yang mungkin disarankan oleh data atau konteks ilmiah dapat digunakan sebagai mesin terbang dengan cara ini. Misalnya, **Gambar 2.13** menunjukkan mesin terbang yang secara bersamaan menampilkan tujuh besaran meteorologi: arah angin, kecepatan angin, tutupan langit, suhu, suhu titik embun, tekanan barometrik, dan kondisi cuaca saat ini. Ketika mesin terbang ini diplot sebagai sebaran yang ditentukan oleh bujur (sumbu horizontal) dan garis lintang (sumbu vertikal), hasilnya adalah peta cuaca mentah yang secara efektif merupakan representasi grafis EDA dari kumpulan data

sembilan dimensi yang menggambarkan distribusi spasial cuaca di waktu tertentu.

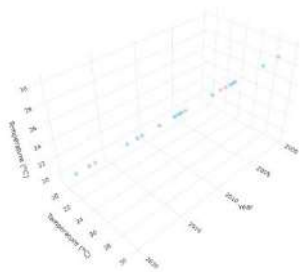
2.3.3 The Rotating Scatterplot

Gambar 2.12 menunjukkan bahwa mencoba memperluas scatterplot dua dimensi menjadi tiga dimensi dengan merendernya sebagai tampilan perspektif umumnya tidak memuaskan. Masalah muncul karena pasangan tiga dimensi dari scatterplot terdiri dari awan titik yang berada dalam volume dan bukan pada bidang, dan memproyeksikan volume tersebut secara geometris ke bidang apa pun menimbulkan ambiguitas tentang jarak yang tegak lurus terhadap bidang tersebut. Salah satu solusi untuk masalah ini adalah dengan menggambar simbol yang lebih besar atau lebih kecil lebih dekat atau lebih jauh dari depan arah proyeksi, dengan meniru perubahan ukuran benda fisik yang terlihat dengan jarak.

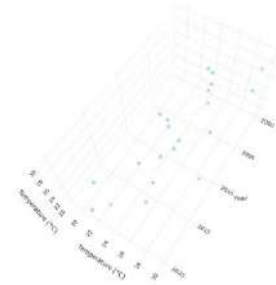
Namun, lebih efektif untuk melihat data tiga dimensi dalam animasi komputer yang dikenal sebagai plot sebar berputar. Pada setiap titik waktu, scatterplot yang berputar adalah proyeksi awan titik tiga dimensi bersama dengan tiga sumbu koordinatnya untuk referensi ke permukaan dua dimensi layar komputer. Tetapi bidang tempat data diproyeksikan dapat diubah dalam waktu, biasanya dengan mouse komputer, untuk memberikan ilusi bahwa kita sedang melihat titik dan sumbunya berputar "di dalam" di sekitar titik asal koordinat tiga dimensi. monitor komputer. Gerakan yang

tampak dapat ditampilkan dengan cukup lancar, dan kontinuitas dalam waktu inilah yang memungkinkan pengertian subjektif dari bentuk data dalam tiga dimensi untuk dikembangkan saat kita mengamati tampilan yang berubah. Faktanya, animasi menggantikan waktu untuk dimensi ketiga yang hilang.

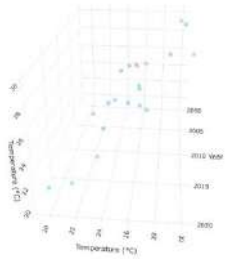
Kekuatan pendekatan ini tidak bisa benar-benar disampaikan dalam bentuk halaman buku yang statis. Namun, gagasan tentang bagaimana ini bekerja dapat diperoleh dari **Gambar 2.14**, yang menunjukkan empat snapshot dari urutan sebar berputar, menggabungkan data suhu, tekanan, dan curah hujan Guayaquil Juni pada Tabel A.3 dengan lima El Niño digunakan selama bertahun-tahun.



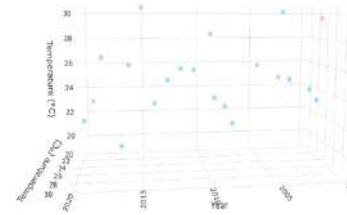
(a)



(b)



(c)



(d)

Gambar 2.14 Empat potret evolusi plot rotasi tiga dimensi dari data Guayaquil Juni pada Tabel A.3, menunjukkan lima tahun El Niño sebagai lingkaran. Sumbu suhu tegak lurus terhadap halaman dan memanjang keluar halaman pada panel (a), dan tiga panel berikutnya menunjukkan perubahan perspektif ketika sumbu suhu diputar pada bidang halaman, ke bawah dan ke kiri. Ilusi visual awan titik yang mengambang di ruang tiga dimensi jauh lebih besar dalam tampilan langsung yang bergerak terus menerus.

ditunjukkan dengan lingkaran terbuka. Awalnya (lihat **Gambar 2.14**a), sumbu suhu diorientasikan keluar dari bidang kertas, yang merupakan sebaran dua dimensi sederhana dari presipitasi versus tekanan. Pada **Gambar 2.14**(b)–(d), sumbu suhu diputar ke dalam bidang halaman, memungkinkan perspektif yang berubah secara bertahap pada penempatan titik relatif satu sama lain dan terhadap proyeksi sumbu koordinat. **Gambar 2.14** hanya menunjukkan rotasi sekitar 90o. Pemeriksaan "langsung" dari data ini dengan rotasi diagram biasanya terdiri dari pemilihan arah awal rotasi (di sini ke bawah, dan ke kiri), biarkan beberapa putaran penuh ke arah itu, dan kemudian mungkin

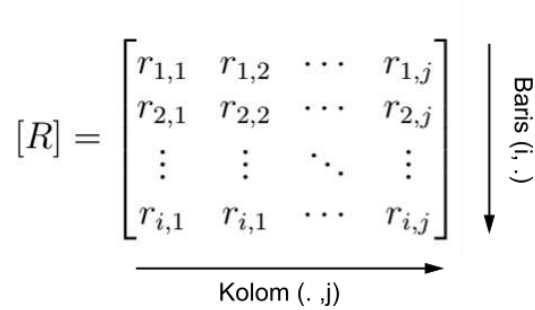
ulangi proses untuk arah putaran lainnya sampai pemahaman tentang bentuk tiga dimensi awan titik berkembang.

2.3.4 Matriks Korelasi

Matriks korelasi adalah alat yang sangat berguna untuk menunjukkan korelasi antara lebih dari dua kumpulan data yang cocok sekaligus. Misalnya, kumpulan data pada Tabel A.1 berisi data yang cocok untuk enam variabel. Koefisien korelasi dapat dihitung untuk masing-masing dari 15 pasangan yang berbeda dari enam variabel tersebut. Secara umum, untuk K variabel, ada $(K)(K-1)/2$ pasangan yang berbeda, dan korelasi antara mereka dapat diatur secara sistematis dalam array persegi dengan baris dan kolom sebanyak variabel data yang cocok yang hubungannya dirangkum menjadi. Setiap entri dalam larik, $r_{i,j}$, diindeks oleh dua indeks i dan j, yang menunjukkan identitas kedua variabel yang korelasinya direpresentasikan. Misalnya, $r_{2,3}$ akan menunjukkan korelasi antara variabel kedua dan ketiga dalam sebuah daftar. Baris dan kolom pada matriks korelasi diberi nomor sesuai sehingga masing-masing korelasi disusun seperti pada **Gambar 2.15**.

Matriks korelasi tidak dirancang untuk analisis data eksplorasi, tetapi sebagai singkatan notasi yang memungkinkan manipulasi matematis dari korelasi dalam kerangka aljabar linier (lihat Bab 9). Sebagai format untuk rangkaian eksplorasi korelasi yang terorganisir, bagian dari

matriks korelasi adalah berlebihan, dan beberapa di antaranya.



Gambar 2.15 Tata letak matriks korelasi, [R]. Korelasi $r_{i,j}$ antara semua pasangan variabel yang mungkin disusun sedemikian rupa sehingga indeks pertama i menunjukkan nomor baris dan indeks kedua j menunjukkan nomor kolom.

hanya tidak informatif. Pertama perhatikan elemen diagonal dari matriks, yang disusun dari sudut kiri atas ke sudut kanan bawah; yaitu, $r_{1,1}, r_{2,2}, r_{3,3}, \ldots, r_{K,K}$. Ini adalah korelasi dari masing-masing variabel terhadap dirinya sendiri dan selalu sama dengan 1. Perhatikan juga bahwa matriks korelasinya simetris. Artinya, korelasi $r_{i,j}$ antara variabel i dan j sama persis dengan korelasi $r_{j,i}$ antara pasangan variabel yang sama, sehingga nilai korelasi di atas dan di bawah diagonal 1 merupakan bayangan cermin satu sama lain. Oleh karena itu, seperti yang telah disebutkan, hanya $(K)(K-1)/2$ dari entri K^2 dalam matriks korelasi yang memberikan informasi unik.

Tabel 1.5 menunjukkan matriks korelasi untuk data pada Tabel A.1. Matriks kiri berisi koefisien korelasi momen

produk Pearson dan matriks kanan berisi koefisien korelasi peringkat Spearman. Seperti yang konsisten dengan praktik umum ketika matriks korelasi digunakan untuk tampilan daripada tujuan komputasi, hanya segitiga bawah dari setiap matriks yang benar-benar dicetak. Elemen diagonal non-informatif dan elemen segitiga atas berlebihan dihilangkan. Hanya $(6)(5)/2 = 15$ nilai korelasi berbeda yang disajikan.

Fitur penting dalam data yang mendasarinya dapat diidentifikasi dengan mempelajari dan membandingkan kedua matriks korelasi ini. Pertama, perhatikan bahwa enam korelasi yang hanya melibatkan variabel suhu memiliki nilai yang sebanding di kedua matriks. Tombak terkuat

Tabel 2.3 Matriks korelasi untuk data pada Tabel A.1. Hanya segitiga bawah dari matriks yang ditunjukkan untuk menghilangkan redudansi dan nilai diagonal yang tidak jelas. Matriks kiri berisi korelasi momen produk Pearson dan matriks kanan berisi korelasi peringkat Spearman.

	Ith. Ppt	Ith. Max	Ith. Ppt	Can. Ppt	Can. Max	Ith. Ppt	Ith. Max	Ith. Ppt	Can. Ppt	Can. Max
Ith. Max	-0					0.32				
Ith. Min	0.29	0.72				0.6	0.76			
Can. Ppt	0.97	0.02	0.27			0.75	0.28	0.55		
Can. Max	-0	0.96	0.76	-0		0.27	0.94	0.75	0.19	
Can. Min	0.22	0.76	0.92	0.19	0.81	0.51	0.79	0.92	0.35	0.78

Korelasi ada antara variabel suhu yang sama di dua lokasi. Korelasi antara suhu maksimum dan minimum pada lokasi yang sama cukup besar tetapi lebih lemah. Korelasi yang melibatkan satu atau kedua variabel curah hujan berbeda

secara signifikan antara dua matriks korelasi. Hanya ada sedikit curah hujan yang sangat besar untuk masing-masing dari dua lokasi, dan ini cenderung mendominasi korelasi Pearson, seperti yang dijelaskan sebelumnya. Oleh karena itu, berdasarkan perbandingan antara matriks korelasi ini, kami menduga bahwa data curah hujan mengandung beberapa outlier, bahkan tanpa mengetahui sifat data atau melihat nomor individu. Korelasi peringkat diharapkan lebih mencerminkan tingkat asosiasi untuk pasangan data yang mempengaruhi satu atau kedua variabel curah hujan. Transformasi variabel curah hujan yang monoton untuk mengurangi kemiringan tidak akan menyebabkan perubahan dalam matriks korelasi Spearman, tetapi akan meningkatkan kesepakatan antara korelasi Pearson dan Spearman.

Dengan sejumlah besar variabel yang terkait dengan korelasinya, jumlah perbandingan berpasangan yang sangat besar bisa sangat banyak, dalam hal ini susunan nilai numerik ini tidak terlalu efektif sebagai perangkat EDA. Namun, area korelasi tertentu dapat diberi warna atau nuansa yang berbeda dan kemudian diplot dalam susunan dua dimensi yang sama dengan korelasi numerik yang menjadi dasarnya, untuk memberikan penilaian visual yang lebih langsung dari pola hubungan.

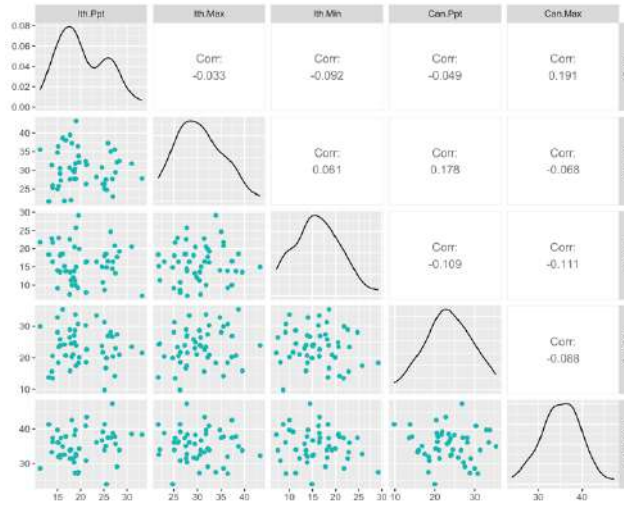
2.3.5 Matriks Scatterplot

Matriks scatterplot adalah perpanjangan grafis dari matriks korelasi. Susunan fisik koefisien korelasi dalam matriks korelasi berguna untuk perbandingan cepat hubungan antara pasangan variabel, tetapi menyaring hubungan tersebut menjadi satu angka seperti koefisien korelasi pasti mengaburkan detail penting. Matriks scatterplot adalah pengaturan plot sebar individu menurut logika yang sama yang menentukan penempatan koefisien korelasi individu dalam matriks korelasi.

Gambar 2.16 adalah matriks scatterplot untuk data Januari 1987, dengan scatterplot disusun dalam pola yang sama dengan matriks korelasi pada Tabel 1.5. Kompleksitas matriks scatterplot dapat membingungkan pada pandangan pertama, tetapi menyajikan banyak informasi tentang perilaku umum data dengan cara yang sangat padat. Pemindaian baris dan kolom presipitasi, misalnya, dengan cepat menunjukkan bahwa hanya ada sedikit presipitasi dalam jumlah besar di masing-masing dari dua lokasi. Melihat secara vertikal di sepanjang kolom presipitasi Ithaca atau secara horizontal di sepanjang baris presipitasi Canandaigua menarik perhatian ke beberapa nilai data terbesar yang tampak berbaris. Sebagian besar titik presipitasi berhubungan dengan jumlah kecil dan karena itu bersarang pada sumbu yang berlawanan. Berfokus pada grafik curah hujan Canandaigua vs. Ithaca, terbukti bahwa kedua lokasi tersebut menerima sebagian besar curah hujan mereka untuk bulan tersebut pada

beberapa hari yang sama. Juga terbukti adanya hubungan presipitasi dengan suhu minimum yang lebih ringan yang terlihat pada tampilan sebelumnya dari data yang sama. Hubungan yang lebih dekat antara variabel suhu maksimum dan maksimum, atau minimum dan minimum, di dua lokasi dibandingkan dengan hubungan antara suhu maksimum dan minimum di satu lokasi juga dapat dilihat dengan jelas.

Matriks scatterplot pada **Gambar 2.5** digambar tanpa elemen diagonal pada posisi yang sesuai dengan korelasi unit variabel dengan variabel itu sendiri dalam korelasi.



Gambar 2.16 Matriks Scatterplot untuk data Januari 1987.

Matriks. Sebuah scatterplot dari variabel apa pun dengan dirinya sendiri akan sama membosankannya, hanya terdiri dari kumpulan titik-titik bujursangkar pada sudut 45o.

Namun, dimungkinkan untuk menggunakan posisi diagonal dalam matriks scatterplot untuk menyajikan informasi univariat yang berguna tentang variabel yang sesuai dengan posisi matriks tersebut. Pilihan sederhana akan menjadi skematis plot masing-masing variabel pada posisi diagonal. Pilihan lain yang berpotensi bermanfaat adalah plot Q-Q (Bagian 4.5.2) untuk setiap variabel, yang secara grafis membandingkan data dengan distribusi referensi; misalnya distribusi Gaussian berbentuk lonceng.

Matriks scatterplot bahkan bisa lebih terbuka saat dibuat dengan perangkat lunak yang memungkinkan untuk "menyikat" titik data di plot terkait. Dengan menyikat, analis dapat memilih titik atau kumpulan titik pada bagan, dan titik yang sesuai pada kumpulan data yang sama juga akan menyala atau dibedakan pada semua bagan lain yang kemudian terlihat. Misalnya, dalam pembuatan **Gambar 2.5**, pembedaan suhu Ithaca pada hari-hari dengan curah hujan terukur dicapai dengan menyikat grafik lain (bagan tidak direproduksi pada **Gambar 2.5**) dengan nilai curah hujan Ithaca. Lingkaran yang diisi pada **Gambar 2.5** dengan demikian mewakili plot pencar suhu yang bergantung pada curah hujan yang tidak nol. Menyikat juga terkadang dapat mengungkapkan hubungan yang mengejutkan dalam data dengan menjaga gerakan menyikat mouse tetap bergerak. "Film" yang dihasilkan dari titik-titik yang disikat di bagan lain yang terlihat secara bersamaan pada dasarnya memungkinkan dimensi waktu tambahan digunakan untuk membedakan hubungan dalam data.

2.3.6 Peta Korelasi

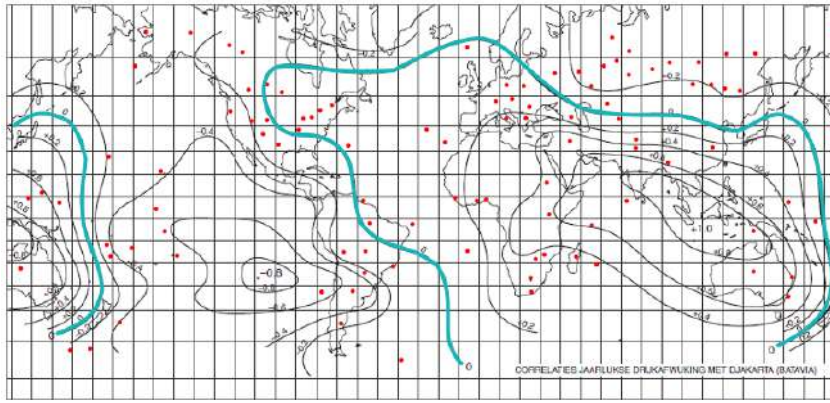
Matriks korelasi seperti pada Tabel 1.5 dapat dipahami dan informatif selama jumlah kuantitas yang diwakili (enam dalam kasus Tabel 1.5) tetap cukup kecil. Ketika jumlah variabel menjadi besar, tidak mungkin untuk dengan mudah memahami setiap nilai atau bahkan menyesuaikan matriks korelasinya pada satu halaman. Penyebab umum dari data atmosfer yang terlalu banyak untuk tampilan yang efektif dalam matriks korelasi atau sebar adalah kebutuhan untuk bekerja dengan data dari sejumlah besar lokasi. Dalam hal ini, susunan geografis lokasi dapat digunakan untuk mengatur informasi korelasi dalam bentuk peta.

Misalnya, pertimbangkan untuk meringkas korelasi antara tekanan permukaan di sekitar 200 lokasi di seluruh dunia. Menurut standar disiplin, ini hanya akan menjadi kumpulan data berukuran sederhana. Namun, banyak tumpukan data tekanan ini akan menghasilkan (200)(199)/2=19.100 pasangan stasiun yang berbeda dan banyak koefisien korelasi. Teknik yang telah berhasil digunakan dalam situasi seperti itu adalah pembuatan serangkaian peta korelasi satu titik.

Gambar 2.17, diambil dari Bjerknes (1969), adalah peta korelasi satu titik untuk data tekanan permukaan tahunan. Peta ini menunjukkan kontur korelasi Pearson antara data tekanan di sekitar 200 lokasi dengan yang ada di Jakarta, Indonesia. Dengan demikian Jakarta adalah "satu titik" dalam peta korelasi satu titik ini. Kuantitas berkontur pada dasarnya

adalah nilai dalam baris (atau kolom) yang sesuai dengan Jakarta dalam matriks korelasi yang sangat besar, yang semuanya mengandung sekitar 19.100 nilai korelasi. Representasi penuh dari matriks korelasi yang besar ini dalam bentuk peta korelasi satu titik akan membutuhkan peta sebanyak stasiun, atau dalam hal ini sekitar 200. Namun, tidak semua peta semenarik **Gambar 2.17**, meskipun peta untuk terdekat stasiun (misalnya B. Darwin, Australia) terlihat sangat mirip.

Jelas, Jakarta berada di bawah +1.0 di peta karena data tekanan di sana berkorelasi sempurna dengan dirinya sendiri. Tidak mengherankan, korelasi tekanan untuk lokasi.



Gambar 2.17 Peta korelasi satu titik tekanan permukaan tahunan di lokasi di seluruh dunia dengan yang ada di Jakarta, Indonesia. Korelasi negatif yang kuat dari -0.8 di Pulau Paskah terkait dengan fenomena El Nino-Osilasi Selatan. Dari Bjerknes (1969).

dekat Jakarta cukup tinggi, dengan penurunan bertahap menuju nol di tempat yang lebih jauh. Pola ini adalah analog spasial dari peluruhan fungsi autokorelasi (temporal) yang diberikan pada **Gambar 2.9**. Hal yang mengejutkan pada **Gambar 2.17** adalah kawasan Pasifik tropis timur yang berpusat di Pulau Paskah, yang korelasinya dengan tekanan Jakarta sangat negatif. Korelasi negatif ini menyiratkan bahwa pada tahun-tahun ketika tekanan rata-rata di Jakarta (dan tempat-tempat terdekat seperti Darwin) tinggi, tekanan di Pasifik timur rendah dan sebaliknya. Pola korelasi ini adalah ekspresi dalam data tekanan permukaan dari fenomena ENSO yang diuraikan sebelumnya dalam bab ini dan merupakan contoh dari apa yang kemudian dikenal sebagai pola telekoneksi. Pada fase hangat ENSO, pusat konveksi Pasifik tropis bergerak ke timur, menciptakan tekanan di bawah rata-rata di dekat Pulau Paskah dan tekanan di atas rata-rata di dekat Jakarta. Saat curah hujan bergeser ke barat selama fase dingin, tekanan rendah di Jakarta dan tinggi di Pulau Paskah.

Tidak semua data korelasi yang tersebar secara geografis menunjukkan pola koneksi seperti yang ditunjukkan pada **Gambar 2.17**. Namun, banyak medan berskala besar, khususnya medan tekanan (atau medan elevasi geopotensial), menunjukkan satu atau lebih pola telekoneksi. Satu perangkat yang digunakan untuk secara bersamaan melihat aspek-aspek dari matriks korelasi besar yang mendasari ini adalah peta telekonektivitas. Untuk menyusun peta telekomunikasi, baris (atau kolom) untuk setiap stasiun atau titik grid dalam

matriks korelasi dicari nilai negatif terbesarnya. Nilai telekonektivitas untuk situs i, T_i, adalah nilai absolut dari yang Korelasi terburuk.

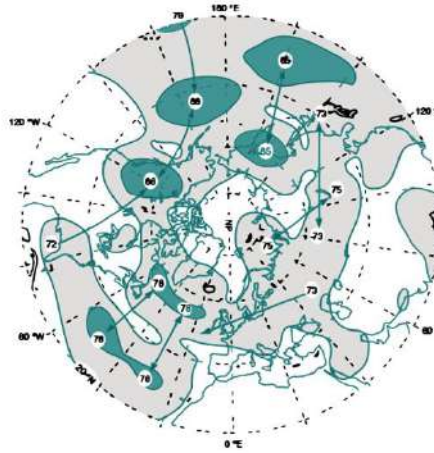
$$T_i = \left| \min_j r_{i,j} \right| \qquad (2.18)$$

Di sini, minimalisasi atas j (indeks kolom untuk [R]) mengimplikasikan bahwa semua korelasi ri,j pada baris ke-i dari [R] dicari untuk nilai terkecil (paling negatif). Misalnya pada **Gambar 2.17**, korelasi negatif terbesar dengan tekanan Jakarta di Pulau Paskah adalah 0,80. Telekonektivitas untuk tekanan permukaan Jakarta akan menjadi 0,80 dan nilai ini akan diplot pada peta telekonektivitas di lokasi Jakarta. Untuk membuat peta telekonektivitas tekanan permukaan penuh, 199 atau lebih baris lain dari matriks korelasi, masing-masing sesuai dengan stasiun yang berbeda, akan diperiksa untuk korelasi negatif terbesar (atau, jika tidak ada yang negatif, maka positif terkecil). dan nilai absolutnya akan diplot pada posisi peta stasiun tersebut.

Gambar 2.18 dari Wallace dan Blackmon (1983) menunjukkan peta telekomunikasi untuk ketinggian 500 mb pada musim dingin di belahan bumi utara. Kepadatan shading menunjukkan besarnya nilai telekonektivitas setiap titik grid. Lokasi telekonektivitas maksimum lokal ditunjukkan oleh posisi angka, dinyatakan sebagai × 100. Panah pada Gambar 1.18 menunjuk dari pusat telekoneksi (yaitu maksima lokal dalam T_i) ke lokasi yang menunjukkan korelasi negatif

maksimum masing-masing. Area yang tidak diarsir mencakup titik jaringan yang telekonektivitasnya relatif rendah. Peta korelasi satu titik untuk lokasi di wilayah yang tidak diarsir ini, analog dengan korelasi waktu pada Gambar 1.9, akan cenderung menunjukkan penurunan bertahap menuju nol dengan jarak, tetapi tanpa turun lebih jauh ke nilai negatif yang besar.

Telah menjadi jelas bahwa sejumlah besar pola telekoneksi ini ada di atmosfer, dan banyak anak panah berkepala dua pada **Gambar 2.18** menunjukkan bahwa secara alami kelompok ini menjadi pola. Yang paling mengesankan adalah pola empat pusat yang membentang dari Pasifik tengah ke Amerika Serikat bagian tenggara dan dikenal sebagai pola Pasifik-Amerika Utara atau PNA. Perhatikan, bagaimanapun, bahwa pola-pola ini di sini muncul dari analisis statistik dan eksplorasi dari sejumlah besar data atmosfer. Jenis pekerjaan ini sebenarnya berakar pada awal abad ke-20 (lihat Brown dan Katz, 1991) dan merupakan contoh yang baik.



Gambar 2.18 Telekonektivitas atau nilai absolut dari korelasi negatif terkuat dari masing-masing peta korelasi satu titik yang banyak digambar pada titik grid dasar, untuk ketinggian 500 mb di musim dingin. Dari Wallace dan Blackmon (1983).

Analisis data eksplorasi dalam ilmu atmosfer, yang mengungkap pola menarik dalam kumpulan data yang sangat besar.

Latihan

1. Bandingkan boxplot dan plot skematis yang mewakili data curah hujan pada Tabel A.3.

2. Gunakan d_λ Hinkley untuk mencari transformasi daya yang sesuai untuk data curah hujan pada Tabel A.2 dengan menggunakan Persamaan 1.3.

3. Buat diagram skematik berdampingan untuk kandidat dan distribusi hasil transformasi akhir yang diperoleh dari Latihan 2.

4. Nyatakan suhu bulan Juni 1951 pada Tabel A.3 sebagai anomali standar.

5. Plot fungsi autokorelasi hingga lag 3 untuk data suhu minimum Ithaca pada Tabel A.1.

BAB 3
DISTRIBUSI PROBABILITAS PARAMETRIK

3.1 Pendahuluan

3.1.1 Distribusi parametrik vs. Empiris

Pada Bab 3, metode untuk menginvestigasi dan merepresentasikan variasi dalam kumpulan data disajikan. Pada intinya, metode ini mengungkapkan bagaimana secara empiris kumpulan data tertentu didistribusikan dalam jangkauannya. Bab ini menyajikan pendekatan peringkasan data yang memperkenalkan bentuk sistematik tertentu, yang disebut distribusi parametrik, untuk merepresentasikan variasi dalam data pokok. Bentuk-bentuk matematis ini sama dengan idealisasi data nyata dan merupakan konstruksi teoretis.

Sebaiknya luangkan waktu sejenak untuk memahami mengapa kami menggunakan kekuatan untuk memaksa data dunia nyata agar sesuai dengan bentuk abstrak. Pertanyaannya layak dipertimbangkan karena distribusi parametrik adalah abstraksi. Mereka mewakili data nyata

hanya kira-kira, meskipun dalam banyak kasus perkiraan sebenarnya bisa sangat baik. Pada dasarnya ada tiga cara yang menggunakan distribusi probabilitas parametrik agar dapat berguna.

Kekompakan. Terutama dengan kumpulan data yang besar, pemrosesan data mentah yang berulang dapat menjadi rumit atau bahkan sangat membatasi. Distribusi parametrik yang pas mengurangi jumlah kuantitas yang diperlukan untuk mengkarakterisasi properti data dari statistik orde-n $\left(x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}\right)$.

Smoothing dan interpolasi. Data nyata tunduk pada variasi sampel yang menyebabkan kesenjangan atau benjolan dalam distribusi empiris. Misalnya, pada **Gambar 1.1** dan **Gambar 2.10**a tidak ada suhu maksimum antara 10°F dan 16°F, walaupun suhu maksimum dalam kisaran ini pasti dapat dan memang terjadi di Ithaca pada bulan Januari. Memaksakan distribusi parametrik pada data ini akan menyajikan kemungkinan suhu yang terjadi serta perkiraan kemungkinan terjadinya.

Ekstrapolasi. Memperkirakan probabilitas untuk kejadian di luar rentang kumpulan data tertentu memerlukan asumsi tentang perilaku yang belum teramati. Mengacu lagi pada **Gambar 2.10**a, probabilitas kumulatif empiris yang terkait dengan suhu terdingin, 9°F, diperkirakan sebesar 0,0213 menggunakan posisi plot Tukey. Probabilitas suhu maksimum sedingin atau sedingin ini dapat diperkirakan

sebesar 0,0213, tetapi tidak ada yang dapat dihitung tentang probabilitas suhu maksimum Januari lebih dingin dari 5°F atau 0°F tanpa pengenalan model probabilitas seperti distribusi parametrik.

Perbedaan telah dibuat antara representasi data empiris dan teoritis, tetapi harus ditekankan bahwa penggunaan distribusi probabilitas parametrik tidak terlepas dari pertimbangan empiris. Secara khusus, sebelum kita mulai memplot data menggunakan fungsi teoretis, kita harus memutuskan antara kandidat bentuk distribusi, menyesuaikan parameter dari distribusi yang dipilih, dan memverifikasi bahwa fungsi yang dihasilkan benar-benar memberikan kesesuaian yang sesuai. Ketiga langkah ini membutuhkan penggunaan data nyata.

3.1.2 Apa itu Distribusi Parametrik?

Distribusi parametrik adalah bentuk sistematik dari abstrak atau bentuk karakteristik. Beberapa dari bentuk sistematik ini muncul secara alami sebagai hasil dari jenis proses penghasil data tertentu, dan jika sesuai, ini adalah kandidat yang masuk akal untuk secara ringkas mewakili variasi dalam kumpulan data. Bahkan ketika tidak ada dasar alami yang kuat di balik pilihan distribusi parametrik tertentu, secara empiris dapat dikatakan bahwa distribusi tersebut merepresentasikan kumpulan data dengan sangat baik.

Sifat spesifik dari distribusi parametrik ditentukan oleh nilai spesifik untuk unit yang disebut parameter distribusi. Misalnya, distribusi Gaussian (atau "normal") memiliki bentuk lonceng simetris yang terkenal sebagai bentuk karakteristiknya. Namun, hanya mengklaim bahwa tumpukan data tertentu, katakanlah suhu rata-rata September di lokasi yang diminati, terwakili dengan baik oleh distribusi Gaussian tidak terlalu memberi tahu tentang sifat data tanpa menentukan distribusi Gaussian mana yang diwakili oleh data tersebut. Faktanya, ada banyak sekali contoh khusus distribusi Gaussian yang sesuai dengan semua kemungkinan nilai dari dua parameter distribusi μ dan a. Tetapi mengetahui, misalnya, bahwa suhu bulanan untuk bulan September diwakili dengan baik oleh distribusi Gaussian dengan μ, 60°F dan $\sigma = 2.5$°F memberikan banyak informasi tentang sifat dan besarnya variasi suhu bulan September lokasi tersebut.

3.1.3 Parameter vs Statistik

Ada risiko kebingungan antara parameter distribusi dan statistik sampel. Parameter distribusi adalah karakteristik abstrak dari distribusi tertentu. Mereka secara ringkas mewakili sifat populasi yang mendasarinya. Sebaliknya, statistik adalah kuantitas apa pun yang dihitung dari sampel data. Biasanya, notasi untuk statistik sampel menggunakan huruf Romawi (i.e., ordinary), dan notasi untuk parameter menggunakan huruf Yunani.

Kebingungan antara parameter dan statistik muncul karena, untuk beberapa distribusi parametrik umum, statistik sampel tertentu merupakan estimator yang baik dari parameter distribusi. Misalnya, standar deviasi sampel s sebuah statistik dapat dirancukan dengan parameter a dari distribusi Gaussian, karena keduanya sering disamakan ketika distribusi Gaussian tertentu paling cocok dengan sampel data. Parameter distribusi ditemukan (dipasang) menggunakan statistik sampel. Namun, proses pemasangan tidak selalu sesederhana dengan distribusi Gaussian, dimana rata-rata sampel ditetapkan sama dengan parameter μ dan standar deviasi sampel sama dengan parameter a.

3.1.4 Distribusi Diskrit vs. Kontinu

Ada dua jenis distribusi parametrik yang sesuai dengan jenis data atau variabel acak yang berbeda. Distribusi diskrit menggambarkan kumpulan acak (yaitu data yang menarik) yang hanya dapat mengambil nilai tertentu. Artinya, nilai yang diterima terbatas atau setidaknya tak terbatas. Misalnya, variabel acak diskrit hanya dapat memiliki nilai 0 atau 1, atau salah satu bilangan bulat nonnegatif, atau salah satu warna merah, kuning, atau biru. Variabel acak kontinu biasanya dapat mengambil nilai apa pun dalam rentang bilangan real tertentu. Misalnya, variabel acak kontinu dapat didefinisikan untuk bilangan real antara 0 dan 1, atau untuk bilangan real nonnegatif, atau untuk beberapa distribusi, bilangan real apa pun.

Tepatnya, menggunakan distribusi kontinu untuk mewakili data yang dapat diamati menyiratkan bahwa pengamatan yang mendasarinya diketahui oleh sejumlah besar orang yang signifikan secara sewenang-wenang. Ini tidak pernah terjadi, tentu saja, tetapi lebih mudah dan tidak terlalu tepat untuk menyatakan sebagai kontinu variabel-variabel yang secara konseptual kontinu tetapi dilaporkan secara terpisah. Suhu dan curah hujan adalah dua contoh nyata yang benar-benar memanjang melintasi sebagian garis bilangan real, tetapi biasanya dilaporkan di Amerika Serikat sebagai kelipatan diskrit 1°F dan 0,01 inci. Sedikit yang hilang dengan memperlakukan pengamatan diskrit ini sebagai sampel dari distribusi bisa dikatakan kontinu.

3.2 Distribusi Diskrit

Ada sejumlah besar distribusi parametrik yang sesuai untuk variabel acak diskrit. Banyak di antaranya dijelaskan dalam volume ensiklopedis oleh Johnson et al. terdaftar. (1992), bersama dengan hasil pada properti mereka. Hanya ada empat di antaranya, distribusi binomial, distribusi geometris, distribusi binomial negatif, dan distribusi Poisson, yang disajikan di sini.

3.2.1 Distribusi Binomial

Distribusi binomial adalah salah satu distribusi parametrik yang paling sederhana dan karena itu sering digunakan dalam buku teks untuk mengilustrasikan penggunaan dan sifat

distribusi parametrik secara lebih umum. Distribusi ini mengacu pada hasil dari situasi dimana pada beberapa kesempatan (kadang-kadang disebut sebagai percobaan) salah satu dari dua peristiwa MECE terjadi. Secara klasik, kedua peristiwa itu disebut sukses dan gagal, tetapi itu adalah sebutan yang sewenang-wenang. Secara lebih umum, salah satu kejadian (misalnya sukses) diberi nomor 1 dan yang lainnya (kegagalan) diberi nomor nol.

Variabel acak *X* adalah jumlah kemunculan peristiwa (ditunjukkan dengan jumlah satu dan nol) dalam sejumlah pengganti tertentu. Jumlah percobaan, N, dapat berupa bilangan bulat positif apa pun, dan variabel X dapat berupa nilai bilangan bulat non-negatif dari 0 (ketika peristiwa yang menarik tidak terjadi sama sekali di N) hingga N (ketika peristiwa terjadi di masing-masing). Distribusi binomial dapat digunakan untuk menghitung probabilitas untuk masing-masing nilai N yang berubah dari X+1 ini ketika dua kondisi terpenuhi: (1) probabilitas terjadinya peristiwa tidak bervariasi dari percobaan ke percobaan (yaitu kemungkinan kejadian adalah stasioner) dan (2) hasil di masing-masing percobaan N tidak tergantung satu sama lain. Kondisi ini jarang terpenuhi, tetapi situasi dunia nyata dapat mendekati ideal ini sehingga distribusi binomial memberikan representasi yang cukup akurat.

Implikasi dari kendala pertama pada probabilitas kejadian yang konstan adalah bahwa peristiwa yang siklus probabilitasnya teratur harus ditangani dengan hati-hati.

Misalnya, peristiwa yang menarik bisa berupa badai petir atau peristiwa petir yang berbahaya di lokasi dimana terdapat variasi harian atau tahunan dalam probabilitas peristiwa tersebut. Dalam kasus seperti itu, periode parsial (misalnya jam atau bulan) dengan probabilitas kejadian yang kira-kira konstan biasanya dianalisis secara terpisah.

Kondisi kedua yang diperlukan untuk penerapan distribusi binomial, yang berkaitan dengan independensi peristiwa, biasanya lebih bermasalah untuk data atmosfer. Misalnya, distribusi binomial biasanya tidak langsung berlaku untuk kejadian harian atau tidak terjadinya presipitasi. Seperti yang ditunjukkan Contoh 2.2, sering kali ada ketergantungan harian yang signifikan antara peristiwa semacam itu. Untuk situasi seperti ini, distribusi binomial dapat digeneralisasikan ke proses stokastik teoretis yang disebut rantai Markov. Di sisi lain, ketergantungan statistik tahun-ke-tahun di atmosfer biasanya cukup lemah sehingga terjadinya atau tidak terjadinya suatu peristiwa dalam periode tahunan berturut-turut dapat dianggap praktis independen. Contoh semacam ini akan disajikan nanti.

Ilustrasi pertama biasa dari distribusi binomial berhubungan dengan lemparan koin. Jika koinnya seimbang, probabilitas kepala atau ekor adalah 0.5 dan tidak berubah dari satu lemparan koin (atau yang setara dengan lemparan koin) ke lemparan berikutnya. Jika $N > 1$ koin dilemparkan pada saat yang sama, hasil dari salah satu koin tidak akan mempengaruhi hasil lainnya. Situasi pelemparan koin dengan

demikian memenuhi semua persyaratan untuk deskripsi menggunakan distribusi binomial: kejadian dikotomis, independen dengan probabilitas konstan.

Pertimbangkan permainan dimana koin yang seimbang dengan $N = 3$ dibalik atau dilempar secara bersamaan, dan kami tertarik pada jumlah kepala X yang dihasilkan. Kemungkinan nilai X adalah 0, 1, 2, dan 3. Keempat nilai ini adalah partisi MECE dari ruang sampel untuk X, dan karenanya probabilitasnya harus berjumlah 1. Dalam contoh sederhana ini, Anda mungkin tidak perlu berpikir secara eksplisit tentang distribusi binomial untuk melihat bahwa probabilitas keempat peristiwa ini masing-masing adalah 1/8, 3/8, 3/8, dan 1/8

Dalam kasus umum, probabilitas untuk masing-masing nilai N+1 dari X diberikan oleh fungsi distribusi probabilitas untuk distribusi binomial

$$\Pr\{X = x\} = \binom{N}{x} p^x (1-p)^{N-x}, \qquad x = 0,1, \dots, N \qquad (3.1)$$

Di sini, sesuai dengan penggunaannya dalam Persamaan 1.1, huruf besar X menunjukkan variabel acak yang nilai pastinya tidak diketahui atau belum diamati. X kecil menunjukkan nilai spesifik dan spesifik yang dapat diambil oleh variabel acak. Distribusi binomial memiliki dua parameter, N dan p. Parameter p adalah probabilitas peristiwa yang akan diinginkan (sukses) yang terjadi pada salah satu percobaan independen N. Diberikan sepasang parameter N dan p,

Persamaan 2.1 adalah fungsi yang menghubungkan setiap nilai diskrit x 0, 1, 2, ..., N dengan probabilitas sehingga $\sum_x \Pr\{X = x\} = 1$. Yaitu mendistribusikan fungsi distribusi Probabilitas probabilitas atas semua peristiwa dalam ruang sampel. Perhatikan bahwa distribusi binomial tidak biasa karena kedua parameternya secara konvensional diwakili oleh huruf Romawi.

Sisi kanan Persamaan 2.1 terdiri dari dua bagian: bagian kombinatorial dan bagian probabilitas. Bagian kombinatorial memberikan jumlah kemungkinan berbeda untuk mewujudkan x hasil sukses dari kumpulan N upaya. Itu diucapkan "N pilih x" dan dihitung menggunakan:

$$\binom{N}{x} = \frac{N!}{x!\,(N-x)!} \tag{3.2}$$

Berdasarkan konvensi $0! = 1$. Misalnya, jika Anda melempar $N = 3$ koin, hanya ada satu cara untuk mendapatkan $x = 3$ kepala: ketiga koin harus muncul kepala. Menggunakan Persamaan 2.2, "tiga pilih tiga" diberikan oleh $3/(3!0!) = (1 \bullet 2 \bullet 3)/(1 \bullet 2 \bullet 3 \bullet 1) = 1$. Ada tiga cara $x = 1$ dapat dicapai: salah satu yang pertama, koin kedua atau ketiga dapat menjadi kepala sedangkan dua koin lainnya adalah ekor; menggunakan Persamaan 2.2 kita mendapatkan $3/(1!2!) = (1 \bullet 2 \bullet 3)/(1 \bullet 1 \bullet 2) = 3$.

Bagian probabilitas dari Persamaan 2.1 mengikuti dari hukum probabilitas multiplikatif untuk peristiwa independen (Persamaan 2.12). Probabilitas urutan tertentu dari tepat x

peristiwa-peristiwa dan $N - x$ bukan peristiwa hanya dikalikan $p - x$ kali dengan dirinya sendiri dan kemudian dikalikan $N = x$ kali dengan $1 - p$ (probabilitas tidak terjadinya). Banyaknya barisan tertentu dari tepat x kejadian dan $N - x$ ketidakterjadian diberikan oleh bagian kombinatorial untuk setiap x, sehingga perkalian dari bagian kombinatorial dan probabilitas pada Persamaan 2.1 memberikan probabilitas kejadian x terjadi, terlepas dari dari penempatan urutan percobaan N.

Contoh 3.1 Distribusi Binomial dan Pembekuan Danau Cayuga, I

Perhatikan data pada Tabel 2.1, yang mencantumkan data tahunan Danau Cayuga di saat danau New York membeku. Danau Cayuga cukup dalam dan hanya membeku setelah cuaca yang sangat dingin dan mendung dalam waktu yang lama. Setiap musim dingin permukaan danau membeku atau tidak. Apakah danau membeku atau tidak pada musim dingin tertentu pada dasarnya tidak tergantung pada apakah danau itu membeku atau tidak selama beberapa tahun terakhir. Asalkan tidak ada perubahan iklim yang signifikan di wilayah tersebut selama dua ratus tahun terakhir, probabilitas danau akan membeku pada tahun tertentu secara efektif konstan selama periode data pada Tabel 2.1. Oleh karena itu, kami berharap distribusi binomial memberikan gambaran statistik yang baik tentang pembekuan danau ini.

Untuk menggunakan distribusi binomial sebagai representasi sifat statistik dari data pembekuan laut, kita perlu menyesuaikan distribusi dengan data. Menyesuaikan distribusi berarti menemukan nilai spesifik untuk parameter distribusi, dalam hal ini p dan N, dimana pada Persamaan 2.1 akan berperilaku mendekati data pada Tabel 2.1. Distribusi binomial unik karena parameter N tergantung pada pertanyaan yang ingin kita tanyakan, bukan pada data itu sendiri. Jika kita ingin menghitung probabilitas bahwa danau akan membeku pada musim dingin mendatang atau dalam satu musim dingin di masa depan, maka N adalah 1. (Kasus khusus Persamaan 2.1 dengan $N = 1$ disebut distribusi Bernoulli.) Jika kita ingin menghitung probabilitas danau membeku, katakanlah, setidaknya sekali selama satu dekade dimasa mendatang, maka $N = 10$.

Tabel 3.1 Data tahunan Danau Cayuga saat membeku.

1796	1904
1816	1912
1856	1934
1875	1961
1884	1979

Parameter binomial p dalam aplikasi ini adalah probabilitas danau akan membeku pada tahun tertentu. Masuk akal untuk memperkirakan probabilitas ini berdasarkan frekuensi relatif peristiwa pembekuan dalam data. Ini adalah tugas yang mudah. Disini, terlepas dari kerumitan kecil karena tidak

mengetahui secara pasti kapan catatan iklim dimulai. Catatan tertulis jelas dimulai paling lambat tahun 1796, tetapi mungkin dimulai beberapa tahun sebelumnya. Misalkan data pada Tabel 2.1 mewakili rekor 220 tahun yang lalu, 10 peristiwa pembekuan yang diamati kemudian menghasilkan estimasi frekuensi relatif untuk binomial p sebesar $\frac{10}{220} = 0.045$.

Kita sekarang dapat menggunakan Persamaan 2.1 untuk memperkirakan probabilitas dari berbagai kejadian yang berkaitan dengan pembekuan danau ini. Jenis peristiwa yang paling sederhana untuk dikerjakan berkaitan dengan danau yang membeku tepat beberapa kali dianggap sebagai x, dalam beberapa tahun tertentu yaitu N. Misalnya, kemungkinan danau akan membeku sekali dalam 10 tahun.

$$\Pr\{X = 1\} = \binom{10}{1} 0.045^1 (1 - 0.045)^{10-1}$$

$$= \frac{10!}{1!\,9!}(0.045)(0.955^9) = 0.30 \qquad (3.3)$$

Contoh 3.2 Distribusi Binomial dan Pembekuan Danau Cayuga, II

Salah satu contoh peristiwa yang sedikit lebih sulit untuk ditangani digambarkan dengan masalah menghitung probabilitas danau akan membeku setidaknya sekali setiap 10 tahun. Dari Persamaan 2.3 jelas bahwa probabilitas ini tidak akan kurang dari 0,30 karena probabilitas untuk kejadian

gabungan diberikan oleh jumlah probabilitas $Pr\{X_1\} + Pr\{X_2\} + \Pr\{X_{10}\}$sil ini mengikuti Persamaan 2.5 dan faktanya bahwa peristiwa ini saling eksklusif: danau tidak dapat membeku tepat satu kali dan tepat dua kali dalam dekade yang sama.

Pendekatan *brute force* untuk masalah ini adalah menghitung semua 10 probabilitas dalam penjumlahan dan kemudian menjumlahkannya. Namun, ini cukup merepotkan dan beberapa pekerjaan dapat dipermudah dengan sedikit lebih memikirkan masalahnya. Pertimbangkan bahwa ruang contoh di sini terdiri dari 11 peristiwa MECE: bahwa danau membeku tepat $0, 1, 2, \cdots$ atau 10 kali dalam satu dekade. Karena probabilitas untuk 11 kejadian ini harus berjumlah 1, akan lebih mudah untuk terus dilanjutkan.

$$\begin{aligned}\Pr\{X \geq 1\} &= 1 - \Pr\{X = 0\} \\ &= 1 - \frac{10!}{0!\,10!}(0.045)^0(0.955)^{10} \\ &= 0.37\end{aligned} \tag{3.4}$$

Perlu dicatat bahwa distribusi binomial dapat diterapkan pada situasi yang secara intrinsik bukan biner dengan mendefinisikan ulang peristiwa secara tepat. Misalnya, suhu secara intrinsik bukanlah biner, atau bahkan secara intrinsik terpisah. Namun, untuk beberapa aplikasi, menarik untuk mempertimbangkan kemungkinan embun beku; yaitu $\Pr\{T \leq 32°\text{F}\}$. Bersama dengan probabilitas peristiwa

komplemennya, maka $\Pr\{T > 32°\text{F}\}$, situasinya adalah satu dalam peristiwa dikotomis dan oleh karena itu dapat menjadi kandidat representasi menggunakan distribusi binomial.

3.2.2 Distribusi Geometrik

Distribusi geometris berkaitan dengan distribusi binomial namun menggambarkan aspek yang berbeda dari situasi konseptual yang sama. Kedua distribusi milik kumpulan studi independen dimana salah satu dari dua peristiwa dikotomis terjadi. Uji coba bersifat independen dalam arti bahwa probabilitas keberhasilan p tidak bergantung pada hasil percobaan sebelumnya, dan urutannya stasioner dalam arti bahwa p tidak berubah sepanjang rangkaian (sebagai akibat dari siklus tahunan misalnya). Agar distribusi geometris dapat diterapkan, kumpulan uji coba harus berurutan.

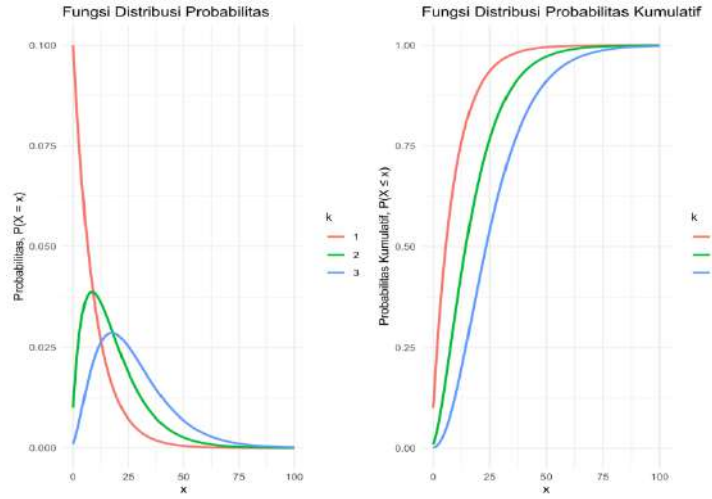
$$\Pr\{X = x\} = p(1-p)^{x-1}, \qquad x = 1,2,\ldots \tag{3.5}$$

Di sini, X dapat mengambil nilai bilangan bulat positif apa pun, karena akan membutuhkan setidaknya satu percobaan untuk mengamati kesuksesan, dan mungkin saja (walaupun sangat kecil kemungkinannya) bahwa kita harus menunggu hasil itu tanpa batas waktu. Persamaan 2.5 dapat dijadikan referensi sebagai penerapan hukum peluang multiplikatif untuk kejadian-kejadian yang saling bebas, karena persamaan ini mengalikan peluang keberhasilan dengan peluang rangkaian kegagalan berturut-turut $x - 1$. Fungsi $k = 1$ pada Gambar 2.1a menunjukkan contoh distribusi probabilitas

geometris untuk probabilitas pembekuan Danau Cayuga $p = 0{,}045$.

3.2.3 Distribusi Binomial Negatif

Distribusi binomial negatif berkaitan erat dengan distribusi geometris, meskipun hubungan ini tidak ditunjukkan dengan namanya namun berasal dari turunannya atau dapat digambarkan sebebagai berikut:



Gambar 3.1 Fungsi distribusi probabilitas (a) dan fungsi distribusi probabilitas kumulatif (b) waktu tunggu $x + k$ tahun agar Danau Cayuga membeku k kali, menggunakan distribusi binomial negatif pada Persamaan 2.6.

Paralel dengan derivasi serupa untuk distribusi binomial. Fungsi distribusi probabilitas untuk distribusi binomial

negatif didefinisikan untuk nilai integer non-negatif dari variabel acak x.

$$\Pr\{X = x\} = \frac{\Gamma(k+x)}{x!\,\Gamma(k)} p^k (1-p)^x, \quad x = 0,1,2,\ldots \tag{3.6}$$

Distribusi tersebut memiliki dua parameter yaitu p, $0 < p < 1$ dan k, $k > 0$. Untuk nilai bilangan bulat dari k, distribusi binomial negatif disebut distribusi Pascal dan memiliki interpretasi yang menarik sebagai perpanjangan dari distribusi geometris dari waktu tunggu untuk keberhasilan pertama dalam urutan percobaan Bernoulli yang independen dengan probabilitas p. Dalam kasus ini, binomial negatif X mengacu pada jumlah kegagalan hingga keberhasilan ke-k, jadi $x + k$ adalah total waktu tunggu yang diperlukan untuk mengamati keberhasilan ke-k.

Notasi $\Gamma(k)$ di sebelah kiri pada Persamaan 2.6 menunjukkan fungsi matematika standar yang dikenal sebagai fungsi gamma, yang ditentukan oleh bilangan bulat tertentu.

$$\Gamma(k) = \int_0^\infty t^{k-1} e^{-t} dt \tag{3.7}$$

Secara umum, fungsi gamma harus dievaluasi secara numerik (Abramowitz dan Stegun, 1984; Press et al., 1986) atau didekati menggunakan nilai tabulasi seperti yang diberikan pada Tabel 2.2. Ini memenuhi hubungan rekursi faktorial

$$\Gamma(k+1) = k\Gamma(k) \qquad (3.8)$$

dimana Tabel 2.2 dapat diperpanjang tanpa batas waktu. Misalnya $\Gamma(3{,}50) = (2{,}50)\Gamma(2{,}50) = (2{,}50)(1{,}50)\Gamma(1{,}50) = (2{,}50)\ (1{,}50)\ (0{,}8862) = 3{,}323$.

Demikian pula $\Gamma(4{,}50) = (3{,}50)\ \Gamma(3{,}50) = (3{,}50)(3{,}323) = 11{,}631$. Fungsi gamma juga dikenal sebagai fungsi faktorial, yang sangat jelas jika asumsinya adalah bilangan bulat (misalnya dalam Persamaan 2.6 ketika k adalah bilangan bulat) ; yaitu, $\Gamma(k+1) = k!$.

Dengan pemahaman fungsi gamma ini, mudah untuk melihat hubungan antara distribusi binomial negatif dengan bilangan bulat k sebagai distribusi *waiting* untuk k sukses dan distribusi geometris (Persamaan 2.5) sebagai distribusi *waiting* untuk sukses pertama, secara berurutan percobaan Bernoulli dengan probabilitas keberhasilan p. Karena X adalah Jumlah kegagalan sebelum keberhasilan ke-k, jumlah upaya untuk mencapai k keberhasilan adalah $x + k$, jadi Persamaan 2.5 dan 4.6 untuk $k = 1$ mengacu pada situasi yang sama. Pembilang pada faktor pertama di ruas kanan Persamaan 2.6 adalah $\Gamma(x+1) = x!$, jadi $x!$ dalam penyebut. Menyadari bahwa $\Gamma(1) = 1$ (lihat Tabel 2.2), Persamaan 2.6 direduksi menjadi Persamaan 2.5, kecuali bahwa Persamaan 2.6 mengacu pada $k = 1$ percobaan tambahan karena juga mengimplikasikan bahwa $k = 1$ keberhasilan pertama.

Tabel 3.2 Nilai fungsi Gamma, $\Gamma(k)$ (Persamaan 2.7), untuk $1.00 \leq k \leq 1.99$.

k	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1	1.000	0.994	0.989	0.984	0.978	0.974	0.969	0.964	0.960	0.956
1	0.951	0.947	0.944	0.940	0.936	0.933	0.930	0.927	0.924	0.921
1	0.918	0.916	0.913	0.911	0.909	0.906	0.904	0.903	0.901	0.899
1	0.898	0.896	0.895	0.893	0.892	0.891	0.890	0.889	0.889	0.888
1	0.887	0.887	0.886	0.886	0.886	0.886	0.886	0.886	0.886	0.886
2	0.886	0.887	0.887	0.888	0.888	0.889	0.890	0.891	0.891	0.892
2	0.894	0.895	0.896	0.897	0.899	0.900	0.902	0.903	0.905	0.907
2	0.909	0.911	0.913	0.915	0.917	0.919	0.921	0.924	0.926	0.929
2	0.931	0.934	0.937	0.940	0.943	0.946	0.949	0.952	0.955	0.958
2	0.962	0.965	0.969	0.972	0.976	0.980	0.984	0.988	0.992	0.996

Contoh 3.3 Distribusi Binomial Negatif dan Pembekuan Cayuga III

Dengan asumsi bahwa pembekuan Danau Cayuga secara statistik terwakili dengan baik oleh serangkaian uji coba Bernoulli tahunan dengan $p = 0.045$, apa yang dapat dikatakan tentang distribusi probabilitas untuk jumlah tahun $x + k$ yang diperlukan untuk menyelesaikan k musim dingin untuk mengamati dimana danau membeku? Seperti disebutkan sebelumnya, probabilitas ini terkait dengan X dalam Persamaan 2.6

Gambar 3.1a menunjukkan tiga distribusi binomial negatif untuk $k = 1$, 2 dan 3, digeser ke kanan sejauh k tahun untuk menunjukkan distribusi waktu tunggu $x + k$. Artinya, titik paling kiri pada ketiga fungsi pada **Gambar 3.1**a semuanya

sesuai dengan $X = 0$ pada Persamaan 2.6. Untuk $k + 1$ fungsi distribusi probabilitas sama dengan distribusi geometris (Persamaan 2.5) dan gambar menunjukkan bahwa probabilitas pembekuan tahun depan dengan konsep Bernoulli $p = 0.045$. Probabilitas tahun ini $x + 1$ akan terjadi Peristiwa pembekuan berikutnya menurun dengan mantap dan cukup cepat sehingga kemungkinan pembekuan pertama terjadi lebih dari satu abad lagi cukup tipis. Tidak mungkin danau membeku selama $k = 2$ kali sebelum tahun depan, jadi probabilitas pertama untuk $k = 2$ yang diplot pada **Gambar 3.1**a adalah $x + k = 2$ tahun, dan probabilitas ini adalah $p^2 = 0.045^2 = 0.0020$. Probabilitas ini meningkat kemungkinan besar menunggu waktu untuk membekukan dua kali pada $x + 2 = 23$ tahun untuk tenggelam lagi, meskipun ada kemungkinan yang tidak dapat diabaikan bahwa danau tidak akan membeku dua kali dalam satu abad. Saat menunggu $k = 3$ membeku, distribusi probabilitas waktu tunggu semakin diratakan dan bergeser lebih jauh ke masa depan.

Cara alternatif untuk melihat distribusi *waiting* ini adalah melalui fungsi distribusi probabilitas kumulatifnya.

$$\Pr\{X \leq x\} = \sum_{X \leq x} \Pr\{X = x\} \tag{3.9}$$

Plot pada **Gambar 3.1**b. Di sini, dalam analogi Persamaan 1.2 untuk fungsi distribusi kumulatif empiris, semua probabilitas untuk waktu tunggu kurang dari atau sama dengan waktu

tunggu yang diinginkan yang dijumlahkan. Untuk $k = 1$ fungsi distribusi kumulatif awalnya meningkat dengan cepat, menunjukkan bahwa kemungkinan pembekuan pertama berakhir dalam beberapa dekade berikutnya cukup tinggi dan hampir pasti bahwa danau akan membeku lagi dalam satu abad (dengan asumsi, bahwa tahunan probabilitas pembekuan p adalah stasioner, sehingga tidak berkurang dari waktu ke waktu, misalnya sebagai akibat dari perubahan iklim). Fungsi-fungsi ini meningkat lebih lambat untuk waktu tunggu pembekuan $k = 2$ dan $k = 3$; dan memberikan probabilitas 0,94 bahwa danau akan membeku setidaknya dua kali dan probabilitas menjadi 0,83 bahwa danau akan membeku setidaknya tiga kali selama abad berikutnya, sekali lagi dengan asumsi iklim tidak bergerak.

Penggunaan distribusi binomial negatif tidak terbatas pada nilai bilangan bulat dari parameter k saja dan jika k diizinkan untuk mengambil nilai positif apa pun, distribusi dapat cocok untuk menggambarkan variasi data yang fleksibel pada hitungan. Misalnya, distribusi binomial negatif (dalam bentuk yang sedikit dimodifikasi) telah digunakan untuk mewakili distribusi periode hari hujan dan kemarau berturut-turut (Wilks 1999a) dengan cara yang lebih fleksibel daripada Persamaan 2.5 karena nilai k bervariasi dari 1 membuat bentuk yang berbeda untuk distribusi, seperti yang ditunjukkan pada **Gambar 3.1**a. Secara umum, nilai parameter yang sesuai harus ditentukan oleh data yang cocok dengan distribusinya. Artinya, nilai spesifik untuk parameter p dan k harus ditentukan yang memungkinkan Persamaan 2.6

untuk melihat sedekat mungkin dengan distribusi empiris dari data yang akan diwakilinya.

Cara termudah untuk menemukan nilai yang sesuai untuk parameter, untuk mengatur distribusi adalah dengan menggunakan metode momen. Untuk menggunakan metode momen, secara matematis kita menyamakan momen sampel dan momen distribusi (atau populasi). Karena ada dua parameter, perlu menggunakan dua momen distribusi untuk mendefinisikannya. Momen pertama adalah rata-rata dan momen kedua adalah varians. Mengenai parameter distribusi, rata-rata distribusi binomial negatif adalah $\mu = \frac{k(1-p)}{p}$ dan variannya adalah $\sigma^2 = \frac{k(1-p)}{p^2}$. Memperkirakan p dan k menggunakan metode momen hanya melibatkan penyetelan ekspresi ini sama dengan momen sampel yang bersesuaian dan menyelesaikan dua persamaan untuk parameter secara bersamaan. Artinya, setiap nilai data x adalah bilangan bulat dari jumlah dan rata-rata dan varians dari x ini dihitung dan disubstitusi ke dalam persamaan berikut:

$$p = \frac{\bar{x}}{s^2} \tag{3.10a}$$

dan

$$k = \frac{\bar{x}^2}{s^2 - \bar{x}} \tag{3.10b}$$

3.2.4 Distribusi Poisson

Distribusi Poisson menggambarkan jumlah peristiwa diskrit yang terjadi dalam rangkaian atau urutan, dan dengan demikian mengacu pada data tentang jumlah yang hanya dapat memiliki nilai bilangan bulat nonnegatif. Biasanya urutan dipahami sebagai temporal; misalnya, terjadinya badai di Samudra Atlantik selama musim badai tertentu. Namun, distribusi Poisson juga dapat diterapkan pada jumlah peristiwa yang terjadi dalam satu atau lebih, jumlah pom bensin di sepanjang jalan raya tertentu atau sebaran hujan es di area kecil.

Peristiwa yang dihitung secara individual bersifat independen dalam arti bahwa peristiwa tersebut tidak bergantung pada apakah atau berapa banyak peristiwa lain yang telah terjadi di tempat lain dalam urutan tersebut. Mengingat frekuensi kemunculan peristiwa, probabilitas sejumlah peristiwa tertentu dalam interval tertentu hanya bergantung pada ukuran interval, biasanya panjang interval waktu dimana peristiwa dihitung. Frekuensi kejadian tidak bergantung pada dimana interval berada dalam waktu atau seberapa sering peristiwa terjadi dalam interval yang tidak tumpang tindih lainnya. Dengan demikian, peristiwa Poisson terjadi secara acak tetapi dengan laju rata-rata yang konstan. Serangkaian peristiwa semacam itu kadang-kadang dikatakan dihasilkan oleh proses Poisson. Seperti pada distribusi binomial, kepatuhan ketat terhadap kondisi independensi ini seringkali sulit dibuktikan dalam data atmosfer, namun distribusi

Poisson masih dapat memberikan representasi yang berguna jika derajat ketergantungannya tidak terlalu kuat. Idealnya, peristiwa Poisson harus sangat jarang terjadi sehingga kemungkinan terjadinya lebih dari satu peristiwa pada waktu yang sama sangat kecil. Cara lain untuk memotivasi distribusi Poisson secara matematis adalah sebagai kasus terbatas dari distribusi binomial karena p cenderung nol dan N cenderung tak terhingga.

Distribusi Poisson memiliki parameter tunggal, μ, yang menentukan pada jumlah rata-rata kemunculannya. Kadang-kadang disebut sebagai intensitas, parameter Poisson memiliki dimensi fisik kejadian per satuan waktu. Fungsi distribusi probabilitas untuk distribusi Poisson adalah

$$\Pr\{X = x\} = \frac{\mu^x e^{-\mu}}{x!}, \qquad x = 0,1,2, \ldots \tag{3.11}$$

mengaitkan probabilitas dengan semua kemungkinan frekuensi X dari nol hingga banyak tak terhingga. Di sini e adalah 2.718 basis logaritma natural. Ruang pengambilan sampel untuk peristiwa Poisson dengan demikian mengandung (dapat dihitung) elemen dalam jumlah tak terbatas. Jelas penjumlahan Persamaan 2.11 untuk x yang berjalan dari nol hingga tak terhingga harus konvergen dan sama dengan 1. Probabilitas yang terkait dengan jumlah hitungan yang sangat besar semakin kecil karena penyebut pada Persamaan 2.11 adalah $x!$.

Untuk menggunakan distribusi Poisson, distribusi tersebut harus disesuaikan dengan sampel data. Sekali lagi, menyesuaikan distribusi berarti menemukan nilai spesifik untuk parameter tunggal μ yang membuat Persamaan 2.11 berperilaku semirip mungkin dengan kumpulan data yang ada. Untuk distribusi Poisson, cara yang baik untuk mengestimasi parameter μ adalah metode momen. Menyesuaikan distribusi Poisson dengan demikian sangat mudah karena satu-satunya parameternya adalah rata-rata jumlah kejadian per satuan waktu, yang dapat langsung diestimasi sebagai rata-rata sampel dari jumlah kejadian per satuan waktu.

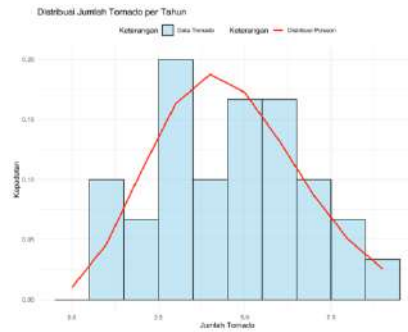
Contoh 3.4 Distribusi Poisson dan Perhitungan Terjadinya Tornado Tahunan Di Negara Bagian New York

Pertimbangkan distribusi Poisson dari jumlah tornado tahunan di Negara Bagian New York untuk tahun 1959–1988 pada Tabel 2.3. Dalam 30 tahun yang dicakup data ini, 138 tornado telah dilaporkan di bagian utara New York. Laju rata-rata atau rata-rata terjadinya tornado adalah 138/30=4,6 tornado/tahun, jadi rata-rata ini adalah estimasi metode sesaat dari intensitas Poisson untuk data ini. Setelah menyesuaikan distribusi dengan mengestimasi nilai parameternya, distribusi Poisson dapat digunakan untuk menghitung probabilitas sejumlah tornado yang akan dilaporkan di New York setiap tahun.

Tabel 3.3 Jumlah Tornado Setiap Tahun di Negara Bagian New York, 1959–1988.

1959	3	1969	7	1979	3
1960	4	1970	4	1980	4
1961	5	1971	5	1981	3
1962	1	1972	6	1982	3
1963	3	1973	6	1983	8
1964	1	1974	6	1984	6
1965	5	1975	3	1985	7
1966	1	1976	7	1986	9
1967	2	1977	5	1987	6
1968	2	1978	8	1988	5

13 pertama dari probabilitas ini (sesuai dengan nol hingga 12 tornado per tahun) disajikan dalam bentuk histogram pada **Gambar 3.2** bersama dengan histogram dari data aktual.



Gambar 3.2 Histogram jumlah tornado yang dilaporkan setiap tahun di Negara Bagian New York selama 1959–1988 (garis putus-putus) dan distribusi Poisson yang sesuai dengan $\mu = 4.6$ tornado/tahun (garis tebal hitam).

Distribusi Poisson memetakan probabilitas dengan lancar (asalkan datanya diskrit) ke hasil yang mungkin, dengan jumlah tornado yang paling mungkin mendekati tingkat rata-rata 4,6. Distribusi data yang ditunjukkan oleh histogram putus-putus mirip dengan distribusi Poisson yang dipasang, tetapi jauh lebih tidak teratur, setidaknya sebagian karena variasi sampel. Misalnya, tampaknya tidak ada alasan yang sah secara fisik mengapa empat tornado per tahun secara signifikan lebih kecil kemungkinannya daripada tiga tornado per tahun, atau mengapa dua tornado per tahun lebih kecil kemungkinannya daripada satu tornado per tahun. Menyesuaikan distribusi Poisson dengan data ini memberikan cara yang masuk akal untuk memuluskan fluktuasi ini, yang diinginkan ketika fluktuasi yang tidak menentu dalam histogram data tidak signifikan secara fisik. Demikian pula, menggunakan distribusi Poisson untuk meringkas data memungkinkan perkiraan kuantitatif probabilitas memiliki nol tornado per tahun atau memiliki lebih dari sembilan tornado per tahun, meskipun jumlah tornado tersebut tidak dilaporkan antara tahun 1959 dan 1988.

3.3 Ekspektasi Statistik

3.3.1 Nilai Harapan dari Variabel Acak

Nilai yang diharapkan dari variabel atau fungsi acak dari variabel acak hanyalah rata-rata tertimbang probabilitas dari variabel atau fungsi tersebut. Rata-rata tertimbang ini disebut sebagai nilai yang diharapkan, meskipun kami tidak selalu percaya bahwa hasil ini mungkin terjadi dalam arti informal dari peristiwa yang "diharapkan". Paradoksnya, bisa terjadi bahwa ekspektasi statistik adalah hasil yang mustahil. Harapan statistik terkait erat dengan distribusi probabilitas karena distribusi memberikan bobot atau fungsi pembobotan untuk rata-rata tertimbang. Kemampuan untuk dengan mudah bekerja dengan ekspektasi statistik dapat menjadi motivator yang kuat untuk menyajikan data menggunakan distribusi parametrik daripada fungsi distribusi empiris.

Paling mudah untuk menganggap ekspektasi sebagai rata-rata dapat ditertimbang probabilitas dalam konteks distribusi probabilitas diskrit seperti distribusi binomial. Secara konvensional, operator ekspektasi dilambangkan dengan $E[\ ]$, sehingga merupakan nilai ekspektasi untuk variabel acak diskrit.

$$E[X] = \sum_{x} x \Pr\{X = x\} \tag{3.12}$$

Notasi ekuivalen <X> = E[X] terkadang digunakan untuk operator ekspektasi. Penjumlahan dalam Persamaan 2.12 diambil alih semua nilai X yang diizinkan. Misalnya, nilai ekspektasi X adalah ketika X mengikuti distribusi binomial

$$E[X] = \sum_{x=0}^{N} x \binom{N}{x} p^x (1-p)^{N-x} \qquad (3.13)$$

Di sini nilai X yang hanya untuk bilangan bulat nonnegatif hingga dan termasuk N, dan setiap suku dalam penjumlahan terdiri dari nilai spesifik variabel x dikalikan dengan probabilitas kemunculannya dari Persamaan 2.1

Ekspektasi E[X] memiliki arti khusus karena merupakan rata-rata dari distribusi X. Sarana distribusi (atau populasi) biasanya dilambangkan dengan simbol μ,. Persamaan 2.12 dapat disederhanakan secara analitik untuk mendapatkan hasil E[X] = Np untuk distribusi binomial. Jadi rata-rata dari setiap distribusi binomial diberikan oleh produk μ = Np. Nilai yang diharapkan untuk keempat distribusi probabilitas diskrit yang dijelaskan dalam Bagian 4.2 tercantum dalam Tabel 2.4 dalam kaitannya dengan parameter distribusi. Data tornado New York pada Tabel 2.3 adalah contoh dari fakta bahwa nilai ekspektasi E[X] = 4.6 tornado tidak layak pada tahun manapun.

3.3.2 Nilai yang diharapkan dari Fungsi Variabel Acak

Akan sangat berguna untuk menghitung ekspektasi atau rata-rata tertimbang probabilitas dari fungsi variabel acak E[g(x)]. Karena ekspektasi adalah operator linier, ekspektasi fungsi variabel acak memiliki sifat berikut:

$$E[c] = c \tag{3.14 a}$$

$$E[c\ g_1(x)] = c\ E[g_1(x)] \tag{3.14 b}$$

$$E\left[\sum_{j=1}^{J} g_j(x)\right] = \sum_{j=1}^{J} E\left[g_j(x)\right] \tag{3.14c}$$

Tabel 3.4 Nilai yang diharapkan (rata-rata) dan varians untuk empat probabilitas diskrit fungsi distribusi yang dijelaskan dalam Bagian 4.2 berkenaan dengan parameter distribusinya.

Distribusi	Fungsi Distribusi Probabilitas	$\mu = E[X]$	$\sigma^2 = Var[X]$
Binomial	Persamaan 2.1	Np	$Np(1-p)$
Geometrik	Persamaan 2.5	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negatif Binomial	Persamaan 2.6	$k(1-p)/p$	$k(1-p)/p$^2
Poisson	Persamaan 2.11	μ	μ

dimana c adalah sembarang konstanta dan $g_j(x)$ sembarang fungsi dari x. Karena konstanta c tidak bergantung pada x, kita memiliki E[c] = $\sum_x$ c Pr{X = x} = c$\sum_x$ Pr{X= x} = c • 1 = c. Persamaan 2.14a dan 4.14b mencerminkan fakta bahwa konstanta dapat difaktorkan keluar saat menghitung

ekspektasi. Persamaan 2.14c menyatakan sifat penting bahwa nilai harapan suatu jumlah sama dengan jumlah nilai harapan individu.

Menggunakan sifat-sifat yang dinyatakan dalam Persamaan 2.14 dapat diilustrasikan dengan ekspektasi fungsi g(x) = (x-μ)². Nilai ekspektasi dari fungsi ini disebut varians dan sering dilambangkan dengan σ^2. Mensubstitusi ke Persamaan 2.12, mengalikan suku-suku dan menerapkan sifat-sifat dalam Persamaan 2.14 memberikan:

$$
\begin{aligned}
Var[X] = E[(X-\mu)^2] &= \sum_x (X-\mu)^2 \Pr\{X = x\} \\
&= \sum_x (x^2 - 2\mu x + \mu^2) \Pr\{X = x\} \\
&= \sum_x x^2 \Pr\{X = x\} - 2\mu \sum_x x \Pr\{X = x\} + \mu^2 \sum_x \Pr\{X = x\} \\
&= E[X^2] - 2\mu E[X] + \mu^2 \\
&= E[X^2] - \mu^2
\end{aligned}
\tag{3.15}
$$

Perhatikan persamaan ruas kanan pertama pada Persamaan 2.15 dengan varians sampel yang diberikan oleh kuadrat Persamaan 1.6. Demikian pula, persamaan akhir dalam Persamaan 2.15 analog dengan bentuk perhitungan varians sampel yang diberikan oleh kuadrat dari Persamaan 1.10. Perhatikan juga bahwa menggabungkan baris pertama Persamaan 2.15 dengan sifat-sifat pada Persamaan 2.14 memberikan.

$$Var[c\ g(x)] = c^2 Var[g(x)] \tag{3.16}$$

Varians untuk empat distribusi diskrit yang dijelaskan dalam Bagian 4.2 tercantum dalam Tabel 2.4

Contoh 3.5 Nilai Harapan dari Fungsi Variabel Acak Binomial

Tabel 2.5 mengilustrasikan perhitungan nilai statistik yang diharapkan untuk distribusi binomial dengan $N = 3$ dan $p = 0.5$. Parameter ini sesuai dengan situasi pelemparan tiga koin sekaligus dan menghitung X jumlah kepala. Kolom pertama menunjukkan hasil yang mungkin dari X, dan kolom kedua menunjukkan probabilitas untuk setiap hasil, dihitung menurut Persamaan 2.1

Tabel 3.5 Probabilitas binomial untuk N = 3 dan p = 0.5 dan konstruksi ekspektasi E[X] dan E[X2] sebagai rata-rata pembobotan probabilitas.

X	$\Pr(X = x)$	$x \cdot \Pr(X = x)$	$x^2 \cdot \Pr(X = x)$
0	0.125	0.000	0.000
1	0.375	0.375	0.375
2	0.375	0.750	1.500
3	0.125	0.375	1.125
		$E[X] = 1.500$	$E[X^2] = 3.000$

Kolom ketiga pada Tabel 2.5 menunjukkan suku-suku individual dalam rata-rata tertimbang probabilitas E[X] = $\sum_x$ [xPr(X=x)]. Penambahan keempat nilai ini menghasilkan E[X]=1.5, seperti yang akan diperoleh dengan mengalikan kedua parameter distribusi μ = Np pada Tabel 2.4

Kolom keempat pada Tabel 2.5 juga menunjukkan konstruksi ekspektasi $E[X^2] = 3.0$. Kita dapat memikirkan ekspektasi ini dalam konteks permainan hipotetis dimana pemain menerima $\$X^2$; Artinya, tidak ada jika nol kepala muncul, \$1 jika satu kepala muncul, \$4 jika dua kepala muncul, dan \$9 jika tiga kepala muncul. Selama banyak putaran permainan ini, pembayaran rata-rata jangka panjang $E[X^2]$ adalah \$3,00. Seseorang yang bersedia membayar lebih dari \$3 untuk memainkan permainan ini akan menjadi bodoh atau cenderung mengambil risiko.

Perhatikan bahwa persamaan tertinggi dalam Persamaan 2.15 untuk distribusi binomial khusus ini dapat diverifikasi menggunakan Tabel 2.5. Di sini $E[X^2] - \mu^2 = 3.0 - (1.5)^2 = 0.75$, konsisten dengan $\text{Var}[X] = Np(1 - p) = 3(0.5)(1 - 0.5) = 0.75$.

3.4 Distribusi Kontinyu

Sebagian besar variabel atmosfer dapat mengambil nilai apa pun dari rangkaian nilai. Suhu, curah hujan, tinggi geopotensial, kecepatan angin, dan besaran lainnya tidak terbatas, setidaknya secara konseptual, pada nilai bilangan bulat dari satuan fisik tempat pengukurannya. Meskipun sistem pengukuran dan pelaporan secara inheren membulatkan pengukuran atmosfer ke nilai diskrit, kumpulan nilai yang dapat dilaporkan cukup besar sehingga sebagian besar variabel masih dapat diperlakukan sebagai kuantitas kontinu.

Ada banyak distribusi parametrik kontinu. Yang paling umum digunakan dalam ilmu atmosfer akan dibahas nanti. Informasi ensiklopedis tentang ini dan banyak distribusi berkelanjutan lainnya dapat ditemukan di Johnson et al. (1994, 1995).

3.4.1 Fungsi Distribusi dan Nilai Harapan

Perhitungan probabilitas untuk variabel kontinu agak berbeda, meskipun mirip dengan variabel acak diskrit. Tidak seperti perhitungan probabilitas untuk distribusi diskrit, yang melibatkan penjumlahan atas fungsi distribusi probabilitas diskontinu (misalnya Persamaan 2.1), perhitungan probabilitas untuk variabel acak kontinu melibatkan integrasi pada fungsi kontinu yang disebut fungsi densitas probabilitas (PDF). PDF terkadang hanya disebut sebagai densitas.

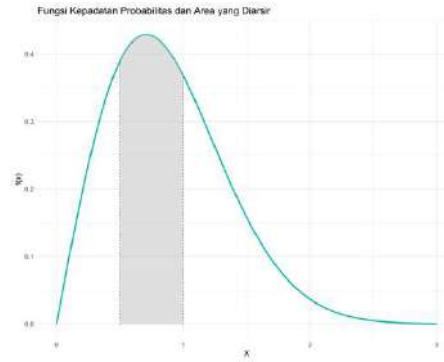
Biasanya fungsi kerapatan probabilitas untuk variabel acak X dilambangkan dengan f(x). Sama seperti jumlah dari fungsi distribusi probabilitas diskrit harus sama dengan 1 untuk semua nilai yang mungkin dari variabel acak, integral dari setiap PDF untuk semua nilai x yang diperbolehkan harus sama dengan 1:

$$\int_x f(x)dx = 1 \tag{3.17}$$

Suatu fungsi tidak bisa menjadi PDF jika tidak memenuhi persamaan ini. Juga, PDF f(x) harus nonnegatif untuk semua nilai x. Tidak ada batas integrasi spesifik yang dimasukkan dalam Persamaan 2.17 karena kepadatan probabilitas yang

berbeda ditentukan pada rentang yang berbeda dari variabel acak (yaitu memiliki dukungan yang berbeda).

Fungsi kepadatan probabilitas adalah analog kontinu dan teoretis dari histogram terkenal (lihat Bagian 3.3.5) dan estimasi kepadatan kernel nonparametrik



Gambar 3.3 Fungsi kepadatan probabilitas hipotetis f(x) untuk variabel acak nonnegatif X. Mengevaluasi f(x) dengan sendirinya tidak berarti dalam hal probabilitas untuk nilai X tertentu. Probabilitas ditemukan dengan mengintegrasikan bagian-bagian dari f(x).

(lihat bagian 3.3.6). Namun, arti PDF sering membingungkan pada awalnya, justru karena analogi dengan histogram. Secara khusus, besarnya fungsi densitas f(x), yang diperoleh ketika dievaluasi untuk nilai tertentu dari variabel acak, dengan sendirinya tidak bermakna dalam pengertian definisi probabilitas. Kebingungan muncul karena yang sering tidak disadari adalah probabilitasnya proporsional dengan luas, bukan tinggi, baik dalam PDF maupun histogram.

Gambar 3.3 menunjukkan PDF hipotetis yang ditentukan pada nilai nonnegatif dari variabel acak X. Fungsi kerapatan probabilitas dapat dievaluasi untuk nilai-nilai tertentu dari variabel acak, katakanlah X = 1, tetapi f(1) saja tidak berarti untuk X dalam hal probabilitas.Faktanya, probabilitas persis X = 1 sangat kecil karena X bervariasi terus-menerus selama segmen perubahan bilangan real. Namun, masuk akal untuk mempertimbangkan dan menghitung probabilitas nilai variabel acak di lingkungan yang tidak terbatas di sekitar X=1. **Gambar 3.3** menunjukkan probabilitas X antara 0,5 dan 1,5 sebagai integral dari PDF antara batas-batas tersebut.

Ide yang terkait dengan PDF adalah fungsi distribusi kumulatif (CDF). CDF adalah fungsi dari variabel acak X yang diberikan oleh integral PDF hingga nilai x tertentu. Jadi, CDF menentukan probabilitas bahwa variabel acak X tidak akan melebihi nilai tertentu. Dengan demikian ini adalah pasangan tetap dari CDF empiris, Persamaan 1.1; dan CDF diskrit, Persamaan 2.9. Secara konvensional, CDF dilambangkan dengan F(x):

$$F(x) = \Pr\{X \leq x\} = \int_{X \leq x} f(x)dx \qquad (3.18)$$

Sekali lagi, batas integrasi tertentu telah dihilangkan dari Persamaan 2.18 untuk menunjukkan bahwa integrasi dilakukan dari nilai minimum X yang diizinkan hingga nilai tertentu x yang merupakan argumen dari fungsi tersebut. Karena nilai F(x) adalah probabilitas, $0 \leq F(x) \leq 1$.

Persamaan 2.18 mengubah nilai tertentu dari variabel acak menjadi probabilitas kumulatif. Nilai variabel acak yang sesuai dengan probabilitas kumulatif yang diberikan diberikan oleh kebalikan dari fungsi distribusi kumulatif

$$F^{-1}(p) = x(F) \tag{3.19}$$

dimana p adalah probabilitas kumulatif. Artinya, Persamaan 2.19 menentukan batas atas integrasi pada Persamaan 2.18 yang memberikan probabilitas kumulatif tertentu p = F(x). Karena invers dari CDF ini memberikan kuantil data yang sesuai dengan probabilitas yang diberikan, Persamaan 2.19 juga dikenal sebagai fungsi kuantil. Bergantung pada distribusi parametrik yang digunakan, dimungkinkan untuk menulis rumus eksplisit untuk CDF atau inversnya.

Ekspektasi statistik juga ditentukan untuk variabel acak kontinu. Seperti pada variabel diskrit, nilai yang diharapkan dari suatu variabel atau fungsi adalah rata-rata tertimbang probabilitas dari variabel atau fungsi tersebut. Karena probabilitas untuk variabel acak kontinu dihitung dengan mengintegrasikan fungsi kerapatannya, nilai yang diharapkan dari fungsi variabel acak diberikan oleh integral.

$$E[g(x)] = \int_x g(x)f(x)dx \tag{3.20}$$

Harapan variabel acak kontinu juga memiliki sifat pada Persamaan 2.14 dan 4.16. Untuk g(x) = x, E[X] = μ adalah rata-rata distribusi yang PDF-nya adalah f(x). Demikian juga,

varian dari variabel kontinu diberikan oleh nilai ekspektasi dari fungsi g(x) = (x – E[X])²

$$Var[X] = E[(x - E[X])^2] = \int_x (x - E[X])^2 f(x) dx \qquad (3.21a)$$

$$= \int_x x^2 f(x) dx - (E[X])^2 = E[X^2] - \mu^2 \qquad (3.21b)$$

Perhatikan bahwa bergantung pada bentuk fungsional tertentu dari f(x), beberapa atau semua integral dalam Persamaan 2.18, 4.20, dan 4.21 mungkin tidak dapat dihitung secara analitik, dan untuk beberapa distribusi integral mungkin tidak ada.

Tabel 3.6 Mencantumkan rata-rata dan varian untuk distribusi yang akan dijelaskan di bagian ini terkait dengan parameter distribusi

Distribusi	PDF	$E[X]$	$Var[X]$
Gaussian	Persamaan 2.23	μ	σ^2
Lognormal[1]	Persamaan 2.30	$\exp[\mu + \sigma^2/2]$	$(\exp[\sigma^2] - 1)\exp[2\mu + \sigma^2]$
Gamma	Persamaan 2.38	$\alpha\beta$	$\alpha\beta^2$
Exponential	Persamaan 2.45	β	β^2
Chi-square	Persamaan 2.47	ν	2ν
Pearson III	Persamaan 2.48	$\zeta + \alpha\beta$	$\alpha\beta^2$
Beta	Persamaan 2.49	$p/(p+q)$	$(pq)/[(p+q)^2(p+q+1)]$
GEV[2]	Persamaan 2.54	$\zeta - \beta[1 - \Gamma(1 - \kappa)]/\kappa$	$\beta^2(\Gamma[1-2\kappa] - \Gamma^2[1-\kappa])/\kappa^2$
Gumbel[3]	Persamaan 2.57	$\zeta + \gamma\beta$	$\beta\pi/\sqrt{6}$
Weibull	Persamaan 2.60	$\beta\Gamma[1 + 1/\alpha]$	$\beta^2(\Gamma[1 + 2/\alpha] - \Gamma^2[1 + 1/\alpha])$
Mixed Exponential	Persamaan 2.66	$w\beta_1 + (1-w)\beta_2$	$w\beta_1^2 + (1-w)\beta_2^2 + w(1-w)(\beta_1 - \beta_2)^2$

3.4.2 Distribusi Gaussian

Distribusi Gaussian memainkan peran sentral dalam statistik klasik dan juga memiliki banyak aplikasi dalam ilmu atmosfer. Hal ini juga kadang-kadang disebut sebagai distribusi normal, meskipun nama itu membawa konotasi yang tidak diinginkan yang dalam beberapa hal bersifat universal atau penyimpangan darinya dalam beberapa hal tidak wajar. PDF-nya adalah kurva berbentuk lonceng yang akrab bahkan bagi orang yang tidak terbiasa dengan statistik.

Luasnya penerapan distribusi Gaussian sebagian besar berasal dari hasil teoretis yang sangat kuat yang dikenal sebagai teorema limit pusat. Secara informal, Teorema Limit Pusat menyatakan bahwa jumlah (atau, ekuivalen, rata-rata aritmatika) dari sekumpulan pengamatan independen memiliki distribusi Gaussian karena ukuran sampel menjadi besar. Ini benar terlepas dari distribusi dari mana pengamatan asli berasal. Pengamatan bahkan tidak harus berasal dari distribusi yang sama! Bahkan, independensi pengamatan tidak benar-benar diperlukan untuk bentuk distribusi yang dihasilkan menjadi Gaussian, yang sangat memperluas penerapan teorema limit pusat untuk data atmosfer.

Apa yang tidak jelas untuk kumpulan data tertentu adalah seberapa besar ukuran sampel yang harus dimiliki oleh teorema limit pusat. Dalam praktiknya, ukuran sampel ini bergantung pada distribusi dari mana penjumlahan diambil.

Jika pengamatan yang dijumlahkan itu sendiri berasal dari distribusi Gaussian, maka jumlah dari sejumlah pengamatan tersebut (termasuk, tentu saja, n=1) juga merupakan distribusi Gaussian. Dengan distribusi yang mendasari tidak terlalu berbeda dengan Gaussian (unimodal dan tidak terlalu asimetris), jumlah dari sejumlah kecil pengamatan hampir mendekati Gaussian. Menyimpulkan suhu harian untuk mendapatkan suhu rata-rata bulanan adalah contoh yang baik dari situasi ini. Pembacaan suhu harian dapat menunjukkan asimetri yang terlihat (misalnya **Gambar 2.5**) tetapi biasanya jauh lebih simetris daripada pembacaan curah hujan harian. Secara konvensional, suhu harian rata-rata didekati sebagai rata-rata suhu maksimum dan minimum harian, sehingga suhu rata-rata bulanan dihitung sebagai:

$$\bar{T} = \frac{1}{30}\sum_{i=1}^{30}\frac{T_{max}(i) + T_{min}(i)}{2} \tag{3.22}$$

selama sebulan dengan 30 hari. Di sini, suhu bulanan rata-rata adalah jumlah dari 60 angka yang diambil dari dua distribusi yang kurang lebih simetris. Mengingat teorema limit pusat, tidak mengherankan bahwa pembacaan suhu bulanan seringkali sangat berhasil diwakili oleh distribusi Gaussian.

Situasi yang kontras adalah total curah hujan bulanan, yang dibangun sebagai jumlah dari, misalnya, 30 nilai curah hujan harian. Meskipun angka yang masuk ke total ini lebih sedikit daripada suhu rata-rata bulanan dalam Persamaan 2.22,

perbedaan yang lebih penting terletak pada distribusi jumlah curah hujan harian yang mendasarinya. Biasanya, sebagian besar nilai curah hujan harian adalah nol, dan sebagian besar jumlah bukan nol kecil. Ini berarti distribusi jumlah curah hujan harian biasanya sangat miring ke kanan. Secara umum, distribusi jumlah dari 30 nilai tersebut juga miring ke kanan, meski tidak terlalu ekstrem. Diagram skematik untuk A1 pada **Gambar 2.3** mengilustrasikan asimetri ini untuk total curah hujan bulan Januari di Ithaca. Akan tetapi, perhatikan bahwa distribusi total curah hujan Ithaca Januari pada **Gambar 2.3** jauh lebih simetris daripada distribusi yang sesuai untuk jumlah curah hujan harian yang mendasari pada Tabel λ=1. Sekalipun penjumlahan 30 nilai harian tidak menghasilkan distribusi Gaussian untuk total bulanan, bentuk distribusi curah hujan bulanan jauh lebih mirip dengan distribusi Gaussian daripada distribusi jumlah curah hujan harian yang sangat miring. Di iklim lembab, distribusi total curah hujan musiman (yaitu, 90 hari) secara bertahap mendekati Gaussian, tetapi bahkan total curah hujan tahunan di lokasi kering dapat menunjukkan kemiringan positif yang signifikan.

PDF untuk distribusi Gaussian adalah

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty \tag{3.23}$$

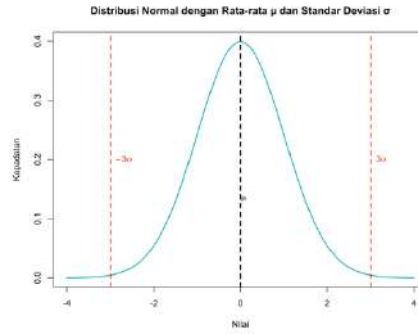
Dua parameter distribusi adalah rata-rata μ dan standar deviasi σ, π adalah konstanta matematika 3.14159 Variabel acak Gaussian didefinisikan secara keseluruhan, jadi Persamaan 2.23 berlaku untuk $-\infty < x < \infty$. Merencanakan Persamaan 2.23 menghasilkan kurva berbentuk lonceng yang terlihat pada **Gambar 2.4**. Gambar ini menunjukkan bahwa rata-rata menempatkan pusat distribusi simetris ini dan standar deviasi mengontrol sejauh mana distribusi menyebar. Hampir semua probabilitas berada dalam $\pm 3\sigma$ dari rata-rata.

Untuk menggunakan distribusi Gaussian untuk mewakili kumpulan data, kedua parameter distribusi harus disesuaikan. Estimasi parameter yang baik untuk distribusi ini mudah diperoleh dengan menggunakan metode momen. Sekali lagi, metode momen tidak lebih dari menyamakan momen sampel dan momen distribusi atau populasi. Momen pertama adalah mean, μ, dan momen kedua adalah varians, σ^2. Oleh karena itu, kita hanya mengestimasi μ sebagai rata-rata sampel (Persamaan 1.2) dan σ sebagai standar deviasi sampel (Persamaan 1.6).

Jika sampel data mengikuti setidaknya kira-kira distribusi Gaussian, estimasi parameter ini menyebabkan Persamaan 2.23 berperilaku mirip dengan data. Kemudian, pada prinsipnya, probabilitas kejadian menarik dapat diperoleh dengan mengintegrasikan Persamaan 2.23. Namun dalam prakteknya, integrasi analitik Persamaan 2.23 tidak mungkin, sehingga rumus untuk CDF, F(x), untuk distribusi Gaussian tidak ada. Sebaliknya, probabilitas Gaussian diperoleh dengan

salah satu dari dua cara. Jika probabilitas diperlukan sebagai bagian dari program komputer, integral dari Persamaan 2.23 dapat didekati secara ekonomis (misalnya Abramowitz dan Stegun 1984) atau dihitung dengan integrasi numerik (misalnya Press et al. 1986) dengan akurasi yang lebih dari cukup. Ketika hanya beberapa probabilitas yang diperlukan, akan lebih mudah untuk menghitungnya dengan tangan menggunakan nilai yang ditabulasikan seperti pada Tabel B.1 di Lampiran B.

Dalam kedua kasus tersebut, transformasi data hampir selalu diperlukan. Ini karena tabel probabilitas dan algoritma Gaussian termasuk dalam standar Gaussian



Gambar 3.4 Fungsi kepadatan probabilitas untuk distribusi Gaussian. Rata-rata μ, menempatkan pusat distribusi simetris ini, dan standar deviasi, σ, mengontrol sejauh mana distribusi menyebar. Hampir semua probabilitas berada dalam $\pm 3\sigma$ dari rata-rata.

Distribusi; yaitu distribusi Gaussian dengan $\mu = 0$ dan $\sigma = 1$. Secara konvensional, variabel acak yang dijelaskan oleh

distribusi Gaussian standar dilambangkan dengan z. Kepadatan probabilitasnya disederhanakan menjadi

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right] \tag{3.24}$$

Simbol $\Phi(z)$ sering digunakan sebagai pengganti $f(z)$ untuk CDF dari distribusi Gaussian standar. Setiap variabel acak Gaussian x dapat diubah menjadi bentuk standar x hanya dengan mengurangkan rata-ratanya dan membaginya dengan standar deviasinya

$$z = \frac{x - \mu}{\sigma} \tag{3.25}$$

Dalam prakteknya, rata-rata dan standar deviasi biasanya perlu diestimasi dari sampel statistik yang sesuai, jadi kami menggunakan

$$z = \frac{x - \bar{x}}{s} \tag{3.26}$$

Perhatikan bahwa semua unit fisik yang mencirikan x dihilangkan dalam transformasi ini, sehingga variabel standar z selalu tidak berdimensi.

Persamaan 2.26 persis sama dengan anomali standar dari Persamaan 1.6. Kumpulan data apa pun dapat diubah dengan mengurangkan rata-rata dan membaginya dengan standar deviasi, dan transformasi ini menghasilkan nilai yang diubah dengan rata-rata sampel nol dan standar deviasi sampel satu.

Namun, data yang ditransformasikan tidak mengikuti distribusi Gaussian kecuali variabel yang tidak ditransformasi mengikutinya. Menggunakan variabel standar dalam Persamaan 2.26 untuk mendapatkan probabilitas diilustrasikan dalam contoh berikut.

Contoh 3.5 Mengevaluasi Probabilitas Gaussian

Pertimbangkan distribusi Gaussian yang ditandai dengan $\mu = 22.2°F$ dan 4.4°F. Nilai-nilai ini telah disesuaikan dengan rangkaian suhu rata-rata bulan Januari di Ithaca. Misalkan kita tertarik untuk memperkirakan probabilitas bahwa bulan Januari yang dipilih secara acak atau bulan Januari mendatang akan memiliki suhu rata-rata minimal 21.4°F, nilai yang diamati pada tahun 1987 (lihat Tabel A.1). Transformasikan suhu ini menggunakan standardisasi pada Persamaan 2.26 z = $(21.4°F - 22.2°F)/4.4°F$= -0.18. Jadi, peluang suhu sedingin atau lebih dingin dari 21.4°F sama dengan peluang nilai z sekecil atau lebih kecil dari -0.18 $\Pr\{X \leq 21.4°F = \Pr\{Z \leq 0.18\}$.

Evaluasi $\Pr\{Z - 0.18\}$ langsung menggunakan Tabel B.1 di Lampiran B, yang berisi probabilitas kumulatif untuk distribusi Gaussian standar. Fungsi distribusi kumulatif ini digunakan secara luas sehingga secara konvensional diberi simbolnya sendiri, $\phi(z)$ bukan $F(z)$ seperti yang diharapkan dari Persamaan 2.18. Melihat melintasi baris berlabel -0.1 ke kolom berlabel 0,08 memberikan probabilitas yang diinginkan sebesar 0,4286. Jelas, ada peluang yang cukup

besar bahwa suhu rata-rata yang sedingin atau sedingin itu akan terjadi di Ithaca pada bulan Januari.

Perhatikan bahwa Tabel B.1 tidak menyertakan baris untuk nilai positif z. Ini tidak diperlukan karena distribusi Gaussian simetris. Ini berarti, misalnya, $\Pr\{Z \geq +0.18\}$ adalah $\Pr\{Z \leq -0.18\}$ karena akan ada luas yang sama di sebelah kiri $z = -0.18$ dan di sebelah kanan $z = +0.18$ di bawah kurva pada **Gambar 3.4**. Jadi, Tabel B.1 dapat digunakan secara lebih umum untuk menaksir probabilitas z > 0 dengan menerapkan relasi

$$\Pr\{Z \leq z\} = 1 - \Pr\{Z \leq -z\} \tag{3.27}$$

yang mengikuti fakta bahwa luas total di bawah kurva dari setiap fungsi kerapatan probabilitas adalah 1 (Persamaan 2.17).

Menggunakan Persamaan 2.27, mudah untuk mengevaluasi $\Pr\{Z \leq \pm 0.18\} = 1 - \ 0.4286 = 0.5714$. Suhu rata-rata Januari di Ithaca, yang sesuai dengan $z = +0.18$, diperoleh dengan membalik Persamaan 2.26

$$x = s\, z + \bar{x} \tag{3.28}$$

Peluangnya adalah 0,5714 bahwa suhu rata-rata bulan Januari di Ithaca tidak akan lebih tinggi dari $(4.4°\text{F})(0.18) + 22.2\ °\text{F} = 23.0°\text{F}$.

Ini hanya sedikit lebih rumit untuk menghitung probabilitas hasil antara dua nilai tertentu, misalnya suhu Ithaca Januari antara 20°F dan 25°F. Karena kejadian $\{T \leq 20°\text{F}\}$ adalah bagian dari kejadian $\{T \leq 25°\text{F}\}$; probabilitas yang diinginkan $\Pr\{20°\text{F} < T \leq 25°\text{F}\}$ diperoleh dengan mengurangkan $\phi(z_{25}) - \phi(z_{20})$. Di sini $z_{25} = \frac{25.0°\text{F} - 22.2°\text{F}}{4.4}°\text{F} = 0.64$ dan $z_{20} = \frac{20.0°\text{F} - 22.2°\text{F}}{4.4}°\text{F} = 0.50$. Jadi (dari Tabel B.1), $\Pr\{20°\text{F} < T \leq 25°\text{F}\} = \phi(z_{25}) - \phi(z_{20}) = 0.73 - 0.309 = 0.430$.

Kadang-kadang juga diperlukan untuk mengevaluasi kebalikan dari CDF Gaussian standar; yaitu, fungsi kuantil Gaussian standar, $\phi^{-1}(p)$. Fungsi ini mengembalikan nilai variabel acak Gaussian standar z yang sesuai dengan probabilitas kumulatif yang ditentukan p. Sekali lagi, tidak mungkin menulis rumus eksplisit untuk fungsi ini, tetapi ϕ^{-1} dapat dievaluasi dalam urutan terbalik menggunakan Tabel B.1. Misalnya, untuk menemukan suhu Ithaca rata-rata di bulan Januari yang menentukan desil bawah (yaitu, 10% terdingin di bulan Januari), isi Tabel B1 akan dicari $\phi(z) = 0.10$. Probabilitas kumulatif ini hampir persis sama $z = -1.28$. Menggunakan Persamaan 2.28, $z = -1.28$ sesuai dengan suhu Januari $(4.4°\text{F})(-1\ 28) + 22.2°\text{F} = 16.6°\text{F}$.

Ketika akurasi tinggi tidak diperlukan untuk probabilitas Gaussian, perkiraan "cukup baik" dari CDF Gaussian standar dapat digunakan

$$\Phi(z) \approx \frac{1}{2}\left[1 \pm \sqrt{1 - \exp\left(\frac{-2z^2}{\pi}\right)}\right] \tag{3.29}$$

Akar positif diambil untuk z > 0 dan akar negatif digunakan untuk z < 0. Kesalahan maksimum yang dihasilkan oleh Persamaan 2.29 adalah sekitar 0,003 (satuan probabilitas) yang terjadi pada $z \pm 1.65$. Persamaan 2.29 dapat dibalik untuk mengaproksimasi fungsi kuantil Gaussian, tetapi aproksimasinya buruk untuk probabilitas ekor (yaitu, ekstrim) yang seringkali paling menarik.

Seperti disebutkan dalam Bagian 3.4.1, salah satu pendekatan untuk menangani data miring adalah dengan melakukan transformasi daya, yang menghasilkan kira-kira distribusi Gaussian. Jika transformasi daya ini adalah log-likelihood (yaitu, $\lambda = 0$ dalam Persamaan 1.3), data (asli, tidak diubah) dikatakan mengikuti distribusi lognormal dengan PDF

$$f(x) = \frac{1}{x\sigma_y\sqrt{2\pi}} \exp\left[-\frac{\left(ln\, x - \mu_y\right)^2}{2\sigma_y^2}\right], \quad x > 0 \tag{3.30}$$

Di sini μ_y dan σ_y masing-masing adalah rata-rata dan standar deviasi dari variabel transformasi y = ln(x). Sebenarnya, nama distribusi logaritma natural agak membingungkan karena

variabel acak x adalah antilog dari variabel y yang mengikuti distribusi Gaussian.

Pemasangan parameter untuk logaritma natural sederhana dan mudah: deviasi rata-rata dan standar dari nilai data yang diubah secara log-likelihood y yaitu μ_y, dan a_y, masing-masing - diestimasi oleh rekan sampel mereka. Hubungan antara parameter ini dalam Persamaan 2.30 dan rata-rata dan varian dari variabel asli X adalah

$$\mu_x = \exp\left[\mu_y + \frac{\sigma_y^2}{2}\right] \tag{3.31a}$$

dan

$$\sigma_x^2 = (\exp\left[\sigma_y^2\right] - 1)\exp[2\mu_y + \sigma_y^2] \tag{3.31b}$$

Probabilitas lognormal dihitung secara sederhana dengan menggunakan variabel transformasi y = ln(x) dan menggunakan perhitungan rutin atau tabel probabilitas distribusi Gaussian. Dalam hal ini, variabel Gaussian standar

$$z = \frac{\ln(x) - \mu_y}{\sigma_y} \tag{3.32}$$

mengikuti distribusi Gaussian dengan $\mu_z = 0$ dan $\sigma_y = 1$

Distribusi lognormal kadang-kadang diasumsikan secara sembarang untuk data yang miring secara positif. Khususnya, lognormal terlalu sering digunakan tanpa memeriksa apakah

transformasi daya lain dapat menghasilkan perilaku Gaussian yang lebih dekat. Secara umum, direkomendasikan agar kandidat transformasi daya lainnya dieksplorasi, seperti yang dibahas di Bagian 3.4.1, sebelum mengasumsikan lognormal untuk kumpulan data tertentu.

Alasan lain, selain kekuatan teorema limit pusat, adalah bahwa distribusi Gaussian digunakan secara luas sehingga dapat dengan mudah digeneralisasikan ke dimensi yang lebih tinggi. Artinya, biasanya mudah untuk merepresentasikan variasi umum dari beberapa variabel Gaussian dengan apa yang disebut distribusi normal multivariat Gaussian atau multivariat. Distribusi ini dibahas lebih detail di Bab 10, karena pengembangan matematis untuk distribusi Gaussian multivariat umumnya memerlukan penggunaan aljabar matriks.

Namun, kasus paling sederhana dari distribusi Gaussian multivariat, yang menggambarkan variasi umum dari dua variabel Gaussian, dapat direpresentasikan tanpa notasi vektor. Distribusi dua variabel ini dikenal sebagai distribusi normal bivariat Gaussian atau bivariat. Kadang-kadang mungkin untuk menggunakan distribusi ini untuk menggambarkan perilaku dua distribusi non-Gaussian jika variabel pertama mengalami transformasi seperti pada Persamaan 1.3. Bahkan, kemampuan untuk menggunakan normal bivariat dapat menjadi motivasi utama untuk menggunakan transformasi tersebut.

Dua variabel yang dipertimbangkan adalah x dan y. Distribusi normal bivariat ditentukan oleh PDF

$$f(x,y) = \frac{1}{2\mu\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right\} \quad (3.33)$$

Persamaan Umum 4.23 dari satu ke dua dimensi, fungsi ini mendefinisikan permukaan di atas bidang x-y daripada kurva di atas sumbu x. Untuk distribusi bivariat kontinu, termasuk normal bivariat, probabilitasnya secara geometris sama dengan volume di bawah permukaan yang ditentukan oleh PDF, jadi dengan analogi dengan Persamaan 2.17 adalah syarat yang diperlukan yang harus dipenuhi oleh setiap PDF bivariat

$$\int_x \int_y f(x,y)dy\,dx = 1\,, \qquad f(x,y) \geq 0 \quad (3.34)$$

Distribusi normal bivariat memiliki lima parameter: dua mean dan standar deviasi untuk variabel x dan y, dan korelasi antara keduanya,ρ. Dua distribusi marjinal untuk variabel x dan y (yaitu fungsi kepadatan probabilitas univariat f(x) dan f(y)) keduanya harus Gaussian. Adalah umum, tetapi tidak dijamin, bahwa distribusi bersama dari dua variabel Gaussian adalah normal bivariat. Dua distribusi marjinal memiliki parameter μ_x, σ_x dan μ_y, σ_y . Pemasangan distribusi normal bivariat sangat mudah. Keempat parameter ini diestimasi

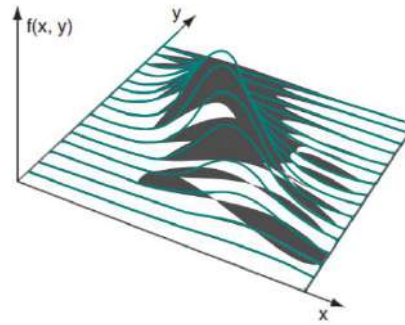
secara terpisah menggunakan contoh pasangannya untuk variabel x dan y, dan parameter ρ diestimasi sebagai korelasi momen produk Pearson antara x dan y, persamaan 1.7.

Gambar 3.5 mengilustrasikan bentuk umum distribusi normal bivariat. Itu berbentuk bukit dalam tiga dimensi, dengan properti yang bergantung pada lima parameter. Fungsi mencapai ketinggian maksimum di atas titik (μ_x, μ_y). Meningkatkan σ_x meregangkan kerapatan dalam arah-x dan meningkatkan σ_y meregangkannya dalam arah-y. Untuk ρ = 0 densitas adalah aksisimetri terhadap titik (μ_x, μ_y), dan kurva dengan tinggi konstan adalah lingkaran konsentris ketika $\sigma_x = \sigma_y$ dan elips sebaliknya. Dengan meningkatnya nilai absolut ρ, fungsi kerapatan diregangkan secara diagonal, dengan garis-garis dengan tinggi konstan menjadi elips yang semakin memanjang. Untuk ρ negatif, orientasi elips ini seperti yang ditunjukkan pada **Gambar 3.5**: nilai x yang lebih besar lebih mungkin dengan nilai y yang lebih kecil, dan nilai x yang lebih kecil lebih mungkin dengan nilai y yang lebih besar. Elips memiliki orientasi yang berlawanan (kemiringan positif) untuk nilai ρ positif.

Probabilitas untuk hasil bersama dari x dan y diberikan, misalnya, dengan integral ganda dari Persamaan 2.33 pada area yang relevan pada bidang

$$\Pr\{(y_1 < Y \leq y_2) \cap (x_1 < X \leq x_2)\} = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x,y) dy\, dx \quad (3.35)$$

Integrasi ini tidak dapat dilakukan secara analitik, dan metode numerik biasanya digunakan dalam praktik. Tabel probabilitas untuk distribusi normal bivariat ada (National Bureau of Standards 1959), tetapi panjang dan tidak praktis. Dimungkinkan untuk menghitung probabilitas untuk daerah berbentuk elips, disebut elips probabilitas, berpusat pada (μ_x, μ_y).



Gambar 3.5 Pandangan perspektif distribusi normal bivariat dengan $\sigma_x = \sigma_y$ dan $= 0.75$. Garis individu yang mewakili punuk distribusi bivariat berbentuk distribusi Gaussian (univariat), menunjukkan bahwa distribusi bersyarat dari x memiliki beberapa nilai y adalah Gaussian sendiri.

Saat menghitung probabilitas untuk wilayah lain, akan lebih mudah untuk bekerja dengan distribusi normal bivariat dalam bentuk standar. Ini adalah perluasan dari distribusi Gaussian univariat standar (Persamaan 2.24) dan dicapai dengan memasukkan variabel x dan y ke dalam transformasi

di Persamaan 2.25. Jadi $\mu_{z_x} = \mu_{z_y} = 0$ dan$\sigma_{z_x} = \sigma_{z_y} = 1$, yang mengarah pada densitas bivariat

$$f(z_x, z_y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{z_x^2 - 2\rho z_x z_y + z_y^2}{2(1-\rho^2)}\right] \qquad (3.36)$$

Properti yang sangat berguna dari distribusi normal bivariat adalah bahwa distribusi bersyarat dari salah satu variabel, mengingat nilai tertentu dari yang lain, adalah Gaussian. Properti ini diilustrasikan secara grafis pada **Gambar 3.5**, dimana garis individual yang menentukan bentuk distribusi dalam tiga dimensi itu sendiri berbentuk Gaussian. Masing-masing menentukan fungsi yang sebanding dengan distribusi bersyarat dari x yang diberi nilai y tertentu. Parameter untuk distribusi Gaussian bersyarat ini dapat dihitung lima parameter normal bivariat. Untuk distribusi x pada nilai y tertentu, fungsi densitas bersyarat Gaussian memiliki parameter f(x|Y = y).

$$\mu_{x|y} = \mu_x + \rho\sigma_x \frac{(y - \mu_y)}{\sigma^y} \qquad (3.37a)$$

dan

$$\sigma_{x|y} = \sigma_x\sqrt{1-\rho^2} \qquad (3.37b)$$

Persamaan 2.37a menghubungkan rata-rata x dengan anomali standar y. Ini menunjukkan bahwa rata-rata bersyarat $\mu_{x|y}$ lebih besar dari rata-rata tak bersyarat μ_x jika y lebih besar

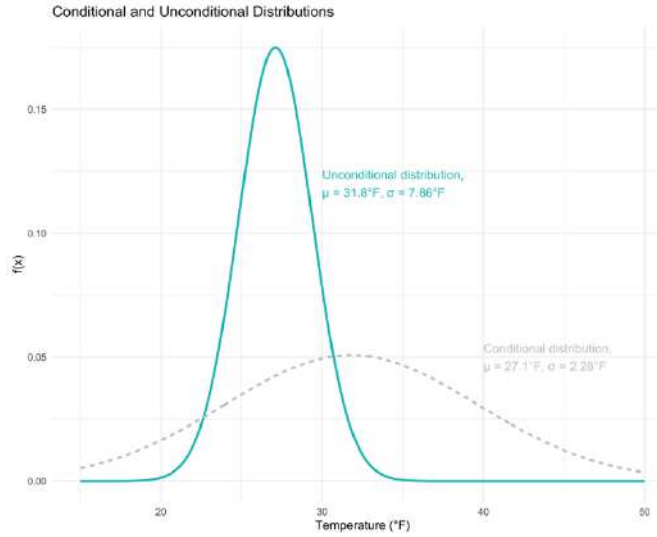
dari rata-ratanya dan ρ positif, atau jika y lebih kecil dari rata-ratanya dan ρ negatif. Jika x dan y tidak berkorelasi, maka mengetahui nilai y tidak memberikan informasi tambahan tentang y dan $\mu_{x|y} = \mu_x$ sejak $\rho = 0$. Persamaan 2.37b menunjukkan bahwa kecuali kedua variabel tidak berkorelasi, $\sigma_{x|y} < \sigma_x$, meskipun ada tanda dari hal. Di sini, pengetahuan tentang y memberikan beberapa informasi tentang x, dan ketidakpastian yang berkurang tentang x tercermin dalam standar deviasi yang lebih kecil. Dalam pengertian ini, ρ^2 sering diinterpretasikan sebagai pecahan varians dalam x yang dijelaskan oleh y.

Contoh 3.6 Distribusi Normal Bivariat dan Probabilitas Bersyarat

Pertimbangkan data suhu maksimum untuk Januari 1987 di Ithaca dan Canandaigua pada Tabel A.1. **Gambar 2.5** menunjukkan bahwa data ini cukup simetris, sehingga berguna untuk memodelkan perilaku sendi mereka sebagai normal bivariat. Sebuah scatterplot dari kedua variabel ini ditunjukkan pada salah satu panel pada **Gambar 2.16**. Maksimum rata-rata Suhu masing-masing adalah 29.87°F dan 31.77°F di Ithaca dan Canandaigua. Simpangan baku sampel yang sesuai adalah 7.71°F dan 7.86°F. Tabel 1.5 menunjukkan korelasi Pearson mereka sebesar 0,957.

Dengan korelasi yang begitu tinggi, mengetahui suhu di satu lokasi seharusnya memberikan informasi yang sangat kuat tentang suhu di lokasi lain. Misalkan diketahui bahwa suhu

maksimum Ithaca adalah 25°F dan informasi probabilitas tentang suhu maksimum Canandaigua diinginkan. Menggunakan Persamaan 2.37a, rata-rata bersyarat untuk distribusi suhu maksimum di Canandaigua, dengan asumsi bahwa suhu maksimum di Ithaca adalah 25°F, adalah 27.1°F-jauh lebih rendah daripada rata-rata tanpa syarat 31.77°F.



Gambar 3.6 Distribusi Gaussian menggambarkan distribusi tak bersyarat untuk suhu maksimum harian Januari di Canandaigua dan distribusi bersyarat dengan asumsi bahwa suhu maksimum di Ithaca adalah 25°F. Korelasi yang tinggi antara suhu maksimum di kedua lokasi berarti bahwa distribusi bersyarat jauh lebih tajam, mencerminkan ketidakpastian yang jauh berkurang.

Menggunakan Persamaan 2.37b, simpangan baku bersyarat adalah 2.28°F. (Ini akan menjadi standar deviasi bersyarat

terlepas dari nilai tertentu dari suhu Ithaca yang dipilih, karena Persamaan 2.37b tidak bergantung pada nilai dari kondisi variabel.) Standar deviasi bersyarat jauh lebih rendah daripada standar deviasi tidak bersyarat karena tingginya korelasi suhu maksimum antara dua lokasi. Seperti yang ditunjukkan pada **Gambar 3.6**, ketidakpastian yang berkurang ini berarti bahwa masing-masing distribusi bersyarat untuk suhu Canandaigua akan jauh lebih tajam mengingat suhu Ithaca daripada distribusi suhu maksimum Canandaigua yang tidak dimodifikasi dan tidak bersyarat.

Dengan menggunakan parameter ini untuk distribusi bersyarat suhu maksimum di Canandaigua, kita dapat menghitung jumlah seperti probabilitas bahwa suhu maksimum di Canandaigua akan berada pada atau di bawah titik beku, mengingat Maksimum Ithaca adalah °F. Variabel standar yang diperlukan adalah z=(32 - 27.1)/2.28 = 2.14, yang sesuai dengan probabilitas 0,984. Sebaliknya, probabilitas klimatologis yang sesuai (tanpa mengetahui suhu maksimum Ithaca) akan dihitung dari z = (32 – 31.8)/7.86 = 0.025, sesuai dengan probabilitas yang jauh lebih rendah 0,510.

3.4.3 Distribusi Gamma

Distribusi statistik dari banyak variabel atmosfer jelas asimetris dan condong ke kanan. Seringkali skewness terjadi ketika ada batas fisik di sebelah kiri yang relatif dekat dengan jangkauan data. Contoh umum adalah jumlah curah hujan

atau kecepatan angin yang dibatasi secara fisik menjadi non-negatif. Meskipun secara matematis dimungkinkan untuk menyesuaikan distribusi Gaussian dalam situasi seperti itu, hasilnya umumnya tidak berguna. Misalnya, data curah hujan Januari 1933–1982 pada Tabel A.2 dapat dicirikan dengan rata-rata sampel 1,96 inci dan standar deviasi sampel sebesar

1,12 inci. Kedua statistik ini cukup untuk menyesuaikan distribusi Gaussian dengan data ini, tetapi penerapan distribusi yang sesuai ini menyebabkan omong kosong. Secara khusus, dengan menggunakan Tabel B.1, kita dapat menghitung probabilitas presipitasi negatif sebagai $\Pr\left\{Z \leq \frac{0.00-1.96}{1.12}\right\} = \Pr\{Z \leq -1.75\} = 0.040$. Probabilitas yang dihitung ini tidak terlalu besar, tetapi juga tidak terlalu kecil. Probabilitas sebenarnya persis nol: tidak mungkin mengamati curah hujan negatif.

Ada berbagai distribusi kontinu yang dibatasi oleh nol dan condong positif. Pilihan umum yang biasanya digunakan untuk merepresentasikan data presipitasi adalah distribusi gamma. Distribusi gamma ditentukan oleh PDF

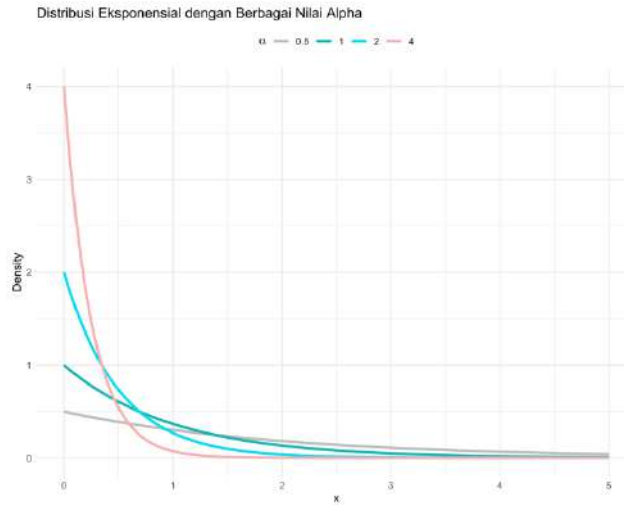
$$f(x) = \frac{\left(\frac{x}{\beta}\right)^{\alpha-1} \exp(-x/\beta)}{\beta\Gamma(\alpha)}, \qquad x, \alpha. \beta > 0 \tag{3.38}$$

Dua parameter distribusi adalah α, parameter bentuk; dan β, parameter skala. Kuantitas $\Gamma(\alpha)$ adalah fungsi gamma yang didefinisikan dalam Persamaan 2.7.

PDF dari distribusi gamma mengambil berbagai bentuk tergantung pada nilai parameter bentuk a. Seperti yang ditunjukkan pada **Gambar 3.7**, untuk α < 1 distribusinya sangat miring ke kanan, dengan $f(x) \to \infty$ sebagai $x \to 0$. Untuk $\alpha < 1$, fungsinya memotong sumbu vertikal pada $\frac{1}{\beta}$ untuk x = 0 (kasus khusus dari distribusi gamma disebut distribusi eksponensial, yang dijelaskan lebih rinci nanti di bagian ini). Untuk α > 1, fungsi densitas distribusi gamma dimulai dari titik asal, f(0) = 0. Nilai yang semakin besar menghasilkan skewness yang lebih kecil dan pergeseran densitas probabilitas ke kanan. Untuk nilai a yang sangat besar (mungkin lebih besar dari 50 hingga 100), distribusi gamma mendekati bentuk Gaussian. Parameter a selalu tanpa dimensi.

Peran parameter skala β efektif untuk merentangkan atau memampatkan (yaitu skala) fungsi kerapatan gamma ke kanan atau kiri, bergantung pada besaran keseluruhan dari nilai data yang diwakili. Perhatikan bahwa variabel acak x dalam Persamaan 2.38 dibagi dengan β di kedua tempat kemunculannya. Parameter skala β memiliki dimensi fisik yang sama dengan x. Karena nilai β yang lebih besar meregangkan distribusi ke kanan, tingginya harus dikurangi untuk memenuhi Persamaan 2.17. Sebaliknya, jika kerapatan digeser ke kiri, ketinggian harus bertambah. Penyesuaian ketinggian ini dicapai dengan β dalam penyebut Persamaan 2.38.

Keserbagunaan bentuk distribusi gamma menjadikannya kandidat yang menarik untuk memplot data curah hujan dan sering digunakan untuk tujuan ini. Namun, ini lebih sulit untuk dikerjakan daripada distribusi Gaussian karena tidak mudah untuk mendapatkan estimasi parameter yang baik dari kumpulan data yang diberikan. Pendekatan termudah (walaupun tentu saja bukan yang terbaik) untuk menyesuaikan distribusi gamma adalah dengan menggunakan metode momen. Tapi ada komplikasi di sini juga, karena dua parameter distribusi gamma tidak sesuai persis dengan momen distribusi, seperti kasus distribusi Gaussian. Rata-rata distribusi gamma diberikan oleh produk $\alpha\beta^2$ dan variannya adalah af32.



Gambar 3.7 Fungsi densitas distribusi gamma untuk empat nilai parameter bentuk α

Menyamakan ekspresi ini dengan contoh yang sesuai kuantitas memberikan satu set dua persamaan dengan dua yang tidak diketahui yang dapat diselesaikan untuk mendapatkan estimasi momen

$$\hat{\alpha} = \bar{x}^2/s^2 \tag{3.39a}$$

dan

$$\hat{\beta} = s^2/\bar{x} \tag{3.39b}$$

Estimator momen untuk distribusi gamma tidak terlalu buruk untuk nilai parameter bentuk yang besar, mungkin $\alpha > 10$, tetapi dapat memberikan hasil yang sangat buruk untuk nilai a yang kecil (Thom 1958; Wilks 1990). Estimator momen dikatakan tidak efisien, dalam arti teknis bahwa mereka tidak memanfaatkan informasi dalam kumpulan data secara optimal. Konsekuensi praktis dari inefisiensi ini adalah bahwa nilai-nilai tertentu dari parameter dalam Persamaan 2.39 adalah variabel yang tidak teratur atau tidak perlu dari sampel data ke sampel data.

Pendekatan yang jauh lebih baik untuk pemasangan parameter untuk distribusi gamma adalah dengan menggunakan metode kemungkinan maksimum. Untuk banyak distribusi, termasuk distribusi gamma, pemasangan kemungkinan maksimum memerlukan prosedur iteratif yang benar-benar hanya praktis dengan komputer. Bagian 4.6 menyajikan metode kemungkinan maksimum untuk

menyesuaikan distribusi parametrik, termasuk distribusi gamma dalam Contoh 4.12.

Ada dua perkiraan penaksir kemungkinan maksimum untuk distribusi gamma yang cukup sederhana untuk dihitung dengan tangan. Keduanya menggunakan statistik sampel

$$D = \ln(\bar{x}) - \frac{1}{n}\sum_{i=1}^{n} \ln(x_i) \tag{3.40}$$

Ini adalah perbedaan antara logaritma natural rata-rata sampel dan rata-rata logaritma data. Demikian pula, statistik sampel D adalah selisih antara logaritma rata-rata aritmetika dan rata-rata geometrik. Perhatikan bahwa rata-rata sampel dan standar deviasi tidak cukup untuk menghitung statistik D karena setiap datum harus digunakan untuk menghitung suku kedua pada Persamaan 2.40.

Yang pertama dari dua perkiraan kemungkinan maksimum untuk distribusi gamma berasal dari Thom (1958). Estimator Thom untuk parameter bentuk adalah

$$\hat{\alpha} = \frac{1 + \sqrt{1 + 4D/3}}{4D} \tag{3.41}$$

setelah itu parameter skala diperoleh dari

$$\hat{\beta} = \frac{\bar{x}}{\hat{\alpha}} \tag{3.42}$$

Pendekatan kedua adalah pendekatan polinomial terhadap parameter bentuk (Greenwood dan Durand 1960). Salah satu dari dua persamaan digunakan

$$\hat{\alpha} = \frac{0.5001 + 0.1639D - 0.0544D^2}{D}, \qquad 0 \leq D \leq 0.5772 \quad (3.43a)$$

Atau

$$\hat{\alpha} = \frac{8.8989 + 9.0599D - 0.0.9775D^2}{17.7972D + 11.9685D^2 + D^3}, \qquad 0.5772 \leq D \leq 17.0 \quad (3.43b)$$

tergantung nilai D. Kemudian parameter skala diestimasi kembali menggunakan Persamaan 2.42.

Seperti distribusi Gaussian, fungsi densitas gamma tidak dapat diintegrasikan secara analitik. Oleh karena itu, probabilitas distribusi gamma harus diperoleh baik dengan menghitung perkiraan ke CDF (yaitu integral dari Persamaan 2.38) atau dari probabilitas tabulasi. Rumus dan rutin komputer untuk tujuan ini dapat ditemukan di Abramowitz dan Stegun (1984) dan Press et al. (1986). Tabel probabilitas distribusi gamma dimasukkan sebagai Tabel B.2 di Lampiran B.

Dalam setiap kasus ini, probabilitas distribusi gamma tersedia untuk distribusi gamma $\beta = 1$ standar. Oleh karena itu, hampir selalu diperlukan untuk mengubah variabel minat X (ditandai dengan distribusi gamma dengan parameter skala arbitrer β) dengan mengubah skala pada variabel standar

$$\xi = x/\beta \tag{3.44}$$

yang mengikuti distribusi gamma dengan β = 1. Variabel gamma default ξ tidak berdimensi. Parameter bentuk α sama untuk x dan ξ. Prosedurnya analog dengan transformasi ke variabel Gaussian z yang dinormalisasi dalam Persamaan 2.26.

Probabilitas kumulatif untuk distribusi gamma standar diberikan oleh fungsi matematika yang dikenal sebagai fungsi gamma parsial $P(a, \xi)$ = Pr{$\Xi \leq \xi$} = $F(\xi)$. Fungsi ini digunakan untuk menghitung probabilitas pada Tabel B.2. Probabilitas kumulatif untuk distribusi gamma standar pada Tabel B.2 disusun terbalik dengan probabilitas Gaussian pada Tabel B.1. Artinya, kuantil (nilai data yang diubah, ξ) dari distribusi disajikan dalam badan tabel, dan probabilitas kumulatif disajikan sebagai judul kolom. Probabilitas yang berbeda diperoleh untuk parameter α yang berbeda yang muncul di kolom pertama.

Contoh 3.7 Mengevaluasi Probabilitas Distribusi Gamma

Ditinjau dari data curah hujan Januari di Ithaca selama 50 tahun 1933–1982 pada Tabel A.2. Curah hujan rata-rata Januari untuk periode ini adalah 1,96 inci dan rata-rata logaritma total curah hujan bulanan adalah 0,5346, memberikan Persamaan 4,40 D = 0.139. Kedua metode Thom (Persamaan 2.41) dan rumus Greenwood dan Durand (Persamaan 2.43a) menghasilkan α = 3.76 dan β = 0.52 inch.

Sebaliknya, penaksir momen (Persamaan 2.39) menghasilkan α = 3 09 dan β = 0.64 inch.

Dengan menggunakan perkiraan kemungkinan maksimum, keanehan total curah hujan Januari 1987 di Ithaca dapat dinilai dengan menggunakan Tabel B.2. Artinya, dengan merepresentasikan variasi klimatologi curah hujan Januari di Ithaca dengan distribusi gamma yang sesuai dengan α = 3.76 dan β = 0.52 inch, probabilitas kumulatifnya bisa sama dengan 3,15 inch (jumlah nilai harian untuk Ithaca pada Tabel A.1).

Pertama, menerapkan Persamaan 2.44, variabel gamma standar ξ = 3.15 inch / 0.52 inch = 6.06. Mengambil α = 3.75 sebagai nilai tabulasi terdekat dengan α = 3.76 yang disesuaikan, dapat dilihat bahwa ξ = 6.06 antara nilai tabulasi adalah untuk F(5.214) = 0.80 dan F(6.354) = 0.90. Interpolasi memberikan F(6.06) = 0.874, menunjukkan bahwa ada sekitar delapan kemungkinan bahwa Januari yang basah atau lebih basah terjadi di Ithaca. Estimasi probabilitas dapat dengan mudah disempurnakan dengan menginterpolasi antara baris untuk α = 3.75 dan β = 3.80 untuk mendapatkan F(6.06) = 0.873, meskipun perhitungan tambahan ini mungkin tidak sepadan dengan usaha.

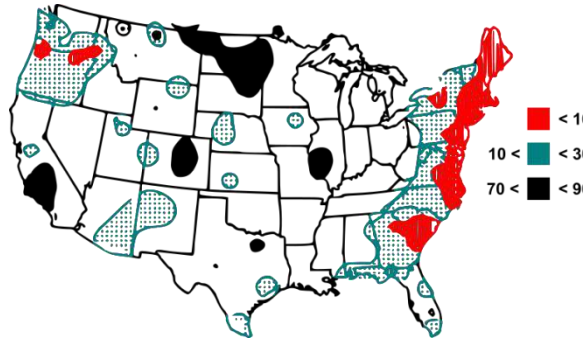
Tabel B.2 juga dapat digunakan untuk membalikkan CDF gamma untuk menemukan nilai curah hujan yang sesuai dengan probabilitas kumulatif tertentu, $\xi = F^{-1}(p)$, yaitu untuk menilai fungsi kuantil. Nilai presipitasi dimensi

kemudian dipulihkan dengan membalikkan transformasi pada Persamaan 2.44. Lihat perkiraan curah hujan rata-rata Ithaca Januari. Ini sesuai dengan nilai ξ yang memenuhi $F(\xi) = 0.50$, yaitu 3,425 berturut-turut untuk α = 3.75 pada Tabel B.2. Jumlah dimensi yang sesuai dari presipitasi diberikan oleh produk $\xi\beta$ = (3.425)(0.52 inch) = 1.78 inch. Sebagai perbandingan, median sampel dari data curah hujan pada Tabel A.2 adalah 1,72 inch. Tidak mengherankan, mediannya adalah kurang dari rata-rata adalah 1,96 inci karena distribusi miring positif. Implikasi (mungkin mengejutkan, tetapi sering diabaikan) dari perbandingan ini adalah karena kemiringan positif dari distribusi curah hujan, curah hujan di bawah normal (yaitu, di bawah rata-rata) lebih mungkin daripada di atas normal.

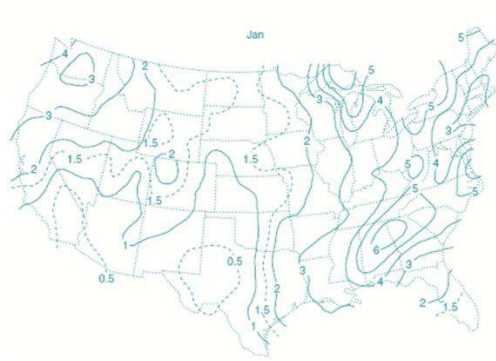
Contoh 3.8 Distribusi Gamma dalam Klimatologi Industri

Distribusi gamma dapat digunakan untuk melaporkan jumlah curah hujan bulanan dan musiman dengan cara yang memungkinkan perbandingan dengan distribusi klimatologi yang berlaku secara lokal. **Gambar 3.8** menunjukkan contoh format curah hujan di Amerika Serikat untuk Januari 1989. Jumlah curah hujan untuk bulan tersebut tidak ditampilkan sebagai kedalaman kumulatif, tetapi sebagai kuantil yang sesuai dengan distribusi gamma klimatologi lokal. Lima kategori dipetakan: kurang dari persentil ke-10 $q_{0.1}$, antara persentil ke-10 dan ke-30 $q_{0.3}$, antara persentil ke-30 dan ke-70 $q_{0.7}$, antara persentil ke-70 dan ke-90 $q_{0.9}$, dan lebih basah dari persentil ke-90 itu

Akan jelas daerah mana yang menerima jauh lebih sedikit, sedikit lebih sedikit curah hujan yang sama, sedikit lebih banyak atau lebih banyak pada Januari 1989 dibandingkan dengan distribusi klimatologis yang mendasarinya. Bentuk distribusi ini sangat berbeda, seperti yang dapat dilihat pada **Gambar 3.9**. Perbandingan dengan **Gambar 3.7** jelas menunjukkan bahwa distribusi curah hujan Januari sangat miring di sebagian besar Barat Daya AS, dan distribusi yang sesuai jauh lebih simetris di sebagian besar Timur dan Pasifik Barat Laut. (Parameter skala yang sesuai dapat diperoleh dari rata-rata curah hujan bulanan dan parameter bentuk dengan$\beta = \frac{\mu}{\alpha}$) Salah satu keuntungan mewakili jumlah curah hujan bulanan dalam kaitannya dengan distribusi gamma klimatologis adalah perbedaan yang sangat kuat dalam bentuk klimatologi presipitasi.



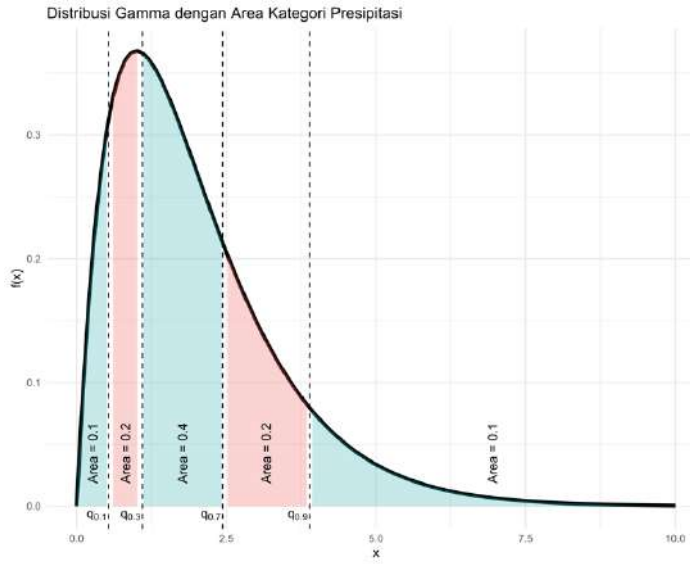
Gambar 3.8 Curah hujan total untuk Januari 1989 di atas Amerika Serikat yang bersebelahan dinyatakan sebagai nilai persentil dari distribusi gamma lokal. Bagian timur dan barat lebih kering dari biasanya, dan bagian tengah negara itu lebih basah. Oleh Arkin (1989).



Gambar 3.9 Parameter bentuk distribusi gamma untuk curah hujan Januari di Amerika Serikat. Distribusi di barat daya sangat miring, dan untuk sebagian besar lokasi di timur jauh lebih simetris. Distribusi dilengkapi dengan data dari 30 tahun 1951-1980. Dari Wilks dan Eggleston (1992).

Jangan bingung membandingkan antar lokasi. Selain itu, menyajikan variasi klimatologis dengan distribusi parametrik memperhalus data klimatologis dan menyederhanakan pembuatan peta dengan meringkas setiap iklim presipitasi hanya menggunakan dua parameter distribusi gamma untuk setiap lokasi, daripada seluruh klimatologi presipitasi mentah untuk Amerika Serikat.

Gambar 3.10 mengilustrasikan definisi persentil menggunakan fungsi densitas probabilitas gamma dengan α = 2. Distribusi dibagi menjadi lima kategori yang sesuai dengan lima tingkat naungan pada **Gambar 3.8** dengan jumlah curah hujan $q_{0.1}$, $q_{0.3}$, $q_{0.7}$ dan $q_{0.9}$ memisahkan wilayah distribusi dengan 10% , 20%, 40%, 20% dan 10%.



Gambar 3.10 Representasi kategori presipitasi pada **Gambar 3.8** dalam bentuk fungsi densitas distribusi gamma dengan α = 2. Hasil yang lebih kering dari persentil ke-10 berada di sebelah kiri $q_{0.1}$. Area dengan curah hujan antara persentil ke-30 dan ke-70 (antara $q_{0.3}$ dan $q_{0.7}$) tidak akan diarsir pada peta. Curah hujan di 10% terbasah dari distribusi klimatologi terletak di sebelah kanan $q_{0.9}$.

Probabilitas atau Seperti yang dapat dilihat pada **Gambar 3.9**, bentuk distribusi pada **Gambar 3.10** adalah karakteristik curah hujan Januari untuk banyak lokasi di Midwest Amerika Serikat dan dataran selatan. Untuk stasiun di Oklahoma timur laut yang melaporkan curah hujan pada Januari 1989 di atas persentil ke-90 pada **Gambar 3.8**, jumlah curah hujan yang sesuai akan lebih besar daripada $q_{0.9}$ yang didefinisikan secara lokal.

Ada dua kasus khusus yang penting dari distribusi gamma yang timbul dari pembatasan tertentu pada parameter α dan β. $\alpha = 1$ mengurangi distribusi gamma menjadi distribusi eksponensial dengan PDF

$$f(x) = \frac{1}{\beta}\exp\left(-\frac{x}{\beta}\right), \quad x \geq 0 \tag{3.45}$$

Bentuk kerapatan ini hanyalah peluruhan eksponensial seperti yang diberikan untuk $\alpha = 1$ pada **Gambar 3.7**. Persamaan 2.45 dapat diintegrasikan secara analitik, sehingga CDF untuk distribusi eksponensial berbentuk tertutup.

$$F(x) = 1 - \exp\left(-\frac{x}{\beta}\right) \tag{3.46}$$

Fungsi kuantil mudah diturunkan dengan menyelesaikan Persamaan 2.46 untuk x (Persamaan 2.8). Karena bentuk distribusi eksponensial ditentukan oleh kendala $\alpha = 1$, biasanya tidak cocok untuk merepresentasikan variasi kuantitas seperti presipitasi, meskipun campuran dari dua distribusi eksponensial (lihat Bagian 4.4.6) dapat merepresentasikan nilai curah hujan harian bukan nol.

Kegunaan penting dari distribusi eksponensial dalam ilmu atmosfer adalah untuk mengkarakterisasi distribusi ukuran tetesan hujan, yang disebut distribusi ukuran tetesan (e.g. Sauvageot 1994). Jika distribusi eksponensial digunakan untuk tujuan ini, itu disebut distribusi Marshall-Palmer dan umumnya dilambangkan dengan N(D), yang menunjukkan

distribusi jumlah tetesan sebagai fungsi dari diameternya. Distribusi ukuran tetesan sangat penting dalam aplikasi radar dimana, misalnya, reflektifitas dihitung sebagai nilai yang diharapkan dari kuantitas yang disebut penampang hamburan balik dalam kaitannya dengan distribusi ukuran tetesan seperti distribusi eksponensial.

Kasus khusus penting kedua dari distribusi gamma adalah distribusi chi-kuadrat (χ^2). Distribusi chi-kuadrat adalah distribusi gamma dengan parameter skala $\beta = 2$. Distribusi chi-kuadrat secara konvensional dinyatakan dalam parameter bilangan bulat yang disebut derajat kebebasan dan dilambangkan dengan v. Secara lebih umum, hubungan dengan distribusi gamma adalah bahwa derajat kebebasan adalah dua kali parameter bentuk distribusi gamma, atau $\alpha = \frac{v}{2}$, memberikan PDF chi-kuadrat

$$f(x) = \frac{x^{(v/2-1)} \exp\left(-\frac{x}{2}\right)}{2^{v/2}\Gamma\left(\frac{v}{2}\right)}, \quad x > 0 \tag{3.47}$$

Karena itu adalah parameter skala gamma yang ditetapkan pada $\beta = 2$ untuk menentukan distribusi chi-kuadrat, Persamaan 2.47 dapat memiliki variasi bentuk yang sama dengan distribusi gamma penuh. Karena tidak ada skala horizontal eksplisit pada **Gambar 3.7**, ini dapat ditafsirkan sebagai menunjukkan kerapatan chi-kuadrat dengan v = 1, 2, 4, dan 8. Distribusi chi-kuadrat diperoleh sebagai distribusi jumlah v kuadrat independen dari variabel Gaussian standar

dan digunakan dalam berbagai cara dalam konteks pengujian statistik (lihat Bab 5). Tabel B.3 mencantumkan kuantil kanan untuk distribusi chi-kuadrat.

Distribusi gamma terkadang juga digeneralisasi menjadi distribusi tiga parameter dengan menggeser PDF ke kiri atau kanan sesuai dengan parameter pergeseran ζ. Distribusi gamma tiga parameter ini juga dikenal sebagai distribusi Pearson tipe III atau hanya Pearson III dan memiliki PDF

$$f(x) = \frac{\left(\frac{x-\zeta}{\beta}\right)\exp\left(-\frac{x-\zeta}{\beta}\right)}{|\beta|\Gamma(\alpha)}, \quad x > \zeta \text{ untuk } \beta > 0, \text{atau } x < \zeta \text{ untuk } \beta < 0 \qquad (3.48)$$

Biasanya parameter skala β adalah positif, sehingga Pearson III menjadi distribusi gamma yang bergeser ke kanan ketika ζ > 0, dengan dukungan x > ζ. Namun, Persamaan 2.48 juga memungkinkan β < 0, dalam hal ini PDF dicerminkan (dan karenanya memiliki ekor kiri panjang dan kemiringan negatif) dan dukungannya adalah x < ζ. Kadang-kadang, secara analogi dengan distribusi lognormal, variabel acak x dalam Persamaan 2.48 telah diubah menjadi log, dalam hal ini distribusi variabel asli [exp(x)] disebut log-Pearson III. Transformasi lain juga dapat digunakan di sini, tetapi asumsi transformasi logaritma tidak sewenang-wenang seperti dalam kasus lognormal. Berbeda dengan bentuk lonceng tetap dari distribusi Gaussian, bentuk distribusi yang sama sekali berbeda dapat diperhitungkan dengan Persamaan 2.48, mirip dengan adaptasi eksponen transformasi λ pada Persamaan 1.3 dengan nilai parameter bentuk α yang berbeda.

3.4.4 Distribusi Beta

Beberapa variabel dibatasi pada segmen garis nyata yang dibatasi pada dua sisi. Variabel tersebut sering dibatasi pada interval $0 \leq x \leq 1$. Contoh variabel penting secara fisik yang tunduk pada batasan ini adalah jumlah awan (diamati sebagai bagian dari langit) dan kelembapan relatif. Variabel penting dan lebih abstrak dari jenis ini adalah probabilitas, dimana distribusi parametrik dapat berguna untuk meringkas frekuensi penggunaan prediksi. probabilitas curah hujan harian. Distribusi parametrik yang biasanya dipilih untuk mendeskripsikan jenis variabel ini adalah distribusi beta.

PDF dari distribusi beta adalah:

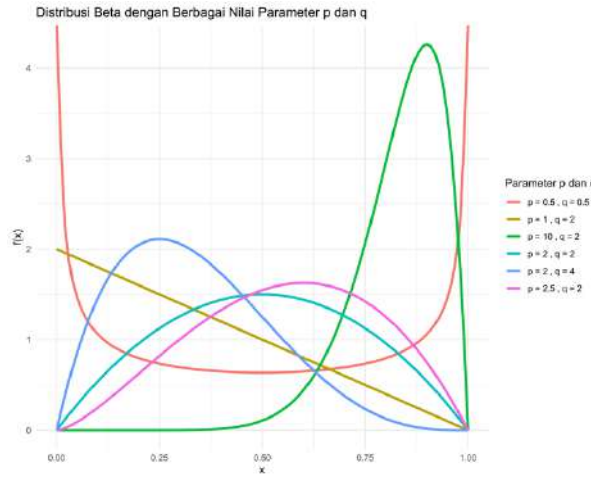
$$f(x) = \left[\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(1)}\right] x^{p-1}(1-x)^{q-1}, \quad 0 \leq x \leq 1, \quad p, q > 0 \quad (3.49)$$

Ini adalah fungsi yang sangat fleksibel yang memiliki banyak bentuk berbeda tergantung pada nilai dari dua parameternya p dan q. **Gambar 3.11** menunjukkan lima di antaranya. Secara umum, untuk $p \leq 1$ probabilitas disatukan mendekati nol (misalnya $p = 25$ dan $q = 2$ atau $p = 1$ dan $q = 2$ pada **Gambar 3.11**), dan untuk $q \leq 1$ probabilitas disatukan mendekati 1. Jika kedua parameter kurang dari satu, maka: Distribusi berbentuk U. Untuk p > 1 dan q > 1, distribusi memiliki mode tunggal (punuk) antara 0 dan 1 (misalnya $p = 2$ dan $q = 4$ atau $p = 10$ dan $q = 2$ pada **Gambar 3.11**), dengan probabilitas bergeser ke kanan untuk lebih banyak

adalah p > q, dan lebih banyak kemungkinan bergeser ke kiri untuk q > p. Distribusi beta dengan $p = q$ simetris. Membalikkan nilai p dan q pada Persamaan 2.49 menghasilkan fungsi densitas yaitu bayangan cermin (dibalik secara horizontal) dari aslinya.

Parameter distribusi beta biasanya dipasang menggunakan metode momen. Menggunakan ekspresi untuk dua distribusi momen pertama yaitu:

$$\mu = p/(p+q) \tag{3.50a}$$



Gambar 3.11 Lima contoh fungsi kerapatan probabilitas untuk distribusi beta. Gambar cermin dari distribusi ini diperoleh dengan membalik parameter p dan q.

dan

$$\sigma^2 = \frac{pq}{(p+q^2)(p+q+1)} \tag{3.50b}$$

dengan penaksir momen

$$\hat{p} = \frac{\bar{x}^2(1-\bar{x})}{s^2} - \bar{x} \tag{3.51a}$$

dan dapat diperoleh

$$\hat{q} = \frac{\hat{p}(1-\bar{x})}{\bar{x}} \tag{3.51b}$$

Kasus khusus yang penting dari distribusi beta adalah distribusi seragam atau persegi panjang dengan p = q = 1 dan PDF f(x) = 1. Distribusi seragam memainkan peran sentral dalam pembuatan angka acak komputer (lihat Bagian 4.7.1).

Penggunaan distribusi beta tidak terbatas pada variabel yang didukung oleh satuan interval [0,1]. Variabel, mis. y dibatasi untuk setiap interval [a,b] dapat diwakili oleh distribusi beta setelah mengalami transformasi

$$x = \frac{y-a}{b-a} \tag{3.52}$$

Dalam hal ini, parameter disesuaikan dengan

$$\bar{x} = \frac{\bar{y}-a}{b-a} \tag{3.53a}$$

dan

$$s_x^2 = \frac{s_y^2}{(b-a)^2} \tag{3.53b}$$

yang kemudian digunakan dalam Persamaan 2.51

Integral kepadatan probabilitas beta tidak ada dalam bentuk tertutup kecuali untuk beberapa kasus khusus, seperti distribusi seragam. Probabilitas dapat diperoleh dengan metode numerik (Abramowitz dan Stegun 1984; Press et al. 1986), dimana CDF untuk distribusi beta dikenal sebagai fungsi beta tidak lengkap $I_x(p,q) = Pr\{0 \leq X \leq x\} = F(x)$. Tabel probabilitas distribusi beta diberikan dalam Epstein (1985) dan Winkler (1972b).

3.4.5 Distribusi Nilai Ekstrim

Statistik nilai ekstrim biasanya dipahami untuk merujuk pada deskripsi perilaku nilai m terbesar. Data ini ekstrem dalam arti sangat besar, dan juga langka menurut definisi. Statistik nilai ekstrim sering menarik perhatian karena proses fisik yang menghasilkan peristiwa ekstrim dan dampak sosial yang diakibatkannya juga besar dan tidak biasa. Sebuah contoh khas dari data nilai ekstrim adalah kumpulan tertinggi tahunan atau tertinggi blok (terbesar dalam blok nilai m) dari nilai curah hujan harian. Dalam setiap n tahun terdapat hari terbasah m = 365 hari dalam setiap tahun dan kumpulan n hari terbasah ini merupakan kumpulan data nilai ekstrim. Tabel 2.7 menunjukkan sampel kecil catatan curah hujan

harian maksimum tahunan untuk Charleston, Carolina Selatan. Untuk setiap n = 20 tahun, tabel menunjukkan jumlah curah hujan terbasah selama m = 365 hari.

Hasil mendasar dari teori statistik nilai ekstrim menyatakan (misalnya Leadbetter et al. 1983; Coles 2001) bahwa pengamatan independen m terbesar dari distribusi tetap akan mengikuti distribusi yang diketahui semakin dekat dengan peningkatan m, terlepas dari (individu, tetap) distribusi dari mana pengamatan berasal. Hasil ini disebut Teorema Tipe Ekstrem dan merupakan analog dalam statistik ekstrem Teorema Limit Pusat untuk distribusi jumlah yang konvergen ke distribusi Gaussian. Teori dan pendekatan sama-sama berlaku untuk distribusi minima ekstrim (pengamatan m terkecil) dengan menganalisis variabel X.

Distribusi yang konvergen dengan distribusi nilai-m-maksimum sampling disebut distribusi nilai ekstrim umum, atau GEV, dengan PDF.

$$f(x) = \frac{1}{\beta}\left[1 + \frac{\kappa(x-\zeta)}{\beta}\right]^{1-1/\kappa} \exp\left\{-\left[1 + \frac{\kappa(x-\zeta)}{\beta}\right]^{-1/\kappa}\right\}, \quad 1 + \frac{\kappa(x-\zeta)}{\beta} > 0 \tag{3.54}$$

Di sini ada tiga parameter: parameter posisi (atau perpindahan) ζ, parameter skala β dan parameter bentuk k. Persamaan 2.54 dapat diintegrasikan secara analitik, menghasilkan CDF

$$F(x) = \exp\left\{-\left[1 + \frac{\kappa(x-\zeta)}{\beta}\right]^{-\frac{1}{\kappa}}\right\} \tag{3.55}$$

Tabel 3.7 Jumlah curah hujan harian maksimum tahunan (inci) di Charleston, Carolina Selatan, 1951–1970

1951	2.01	1956	3.86	1961	3.48	1966	4.58
1952	3.52	1957	1.16	1962	4.60	1967	6.23
1953	2.61	1958	4.20	1963	5.20	1968	2.67
1954	3.89	1959	4.48	1964	4.93	1969	5.24
1955	1.82	1960	4.51	1965	3.50	1970	3.00

dan CDF ini dapat dibalik untuk mendapatkan rumus eksplisit untuk fungsi kuantil

$$F^{-1}(p) = \zeta + \frac{\beta}{\kappa}\{[-\ln(p)]^{\kappa} - 1\} \tag{3.56}$$

Persamaan 2.54 sampai 4.56 sering ditulis dengan tanda kebalikan dari parameter bentuk k, terutama dalam literatur hidrologi.

Karena momen GEV (lihat Tabel 2.6) melibatkan fungsi gamma, pendugaan parameter GEV menggunakan metode momen tidak lebih mudah daripada metode alternatif yang memberikan hasil yang lebih akurat. Distribusi biasanya dipasang baik menggunakan metode kemungkinan maksimum (lihat Bagian 4.6) atau metode yang dikenal sebagai momen-L (Hosking 1990; Stedinger et al. 1993), yang biasa digunakan dalam aplikasi hidrologi. Pemasangan

momen-L cenderung lebih disukai untuk sampel data kecil (Hosking 1990). Metode kemungkinan maksimum dapat dengan mudah diadaptasi untuk menyertakan efek kovariat atau pengaruh tambahan; misalnya, kemungkinan bahwa satu atau lebih dari parameter distribusi dapat menjadi tren karena perubahan iklim (Katz et al. 2002; Smith 1989; Zhang et al. 2004). Untuk ukuran sampel sedang dan besar, hasil dari kedua metode estimasi parameter biasanya serupa. Dengan menggunakan data pada Tabel 2.7, estimasi kemungkinan maksimum untuk parameter GEV ζ = 3.50, β = 1.11, dan k = -0.29; dan perkiraan momen-L yang sesuai adalah ζ = 3.49, β = 1.18, dan k = 0.32.

Bergantung pada nilai parameter k, tiga kasus khusus GEV dikenali. Batas Persamaan 2.54 saat k mendekati nol menghasilkan PDF.

$$f(x) = \frac{1}{\beta}\exp\left\{-\exp\left[-\frac{(x-\zeta)}{\beta}\right] - \frac{(x-\zeta)}{\beta}\right\} \qquad (3.57)$$

dikenal sebagai distribusi Gumbel atau Fisher-Tippett Tipe I. Distribusi Gumbel adalah bentuk terbatas dari GEV untuk data ekstrem yang diambil secara independen dari distribusi dengan perilaku baik (yaitu, eksponensial) seperti Gaussian dan Gamma. Distribusi Gumbel sangat umum digunakan untuk mewakili statistik ekstrim yang kadang-kadang disebut sebagai distribusi nilai ekstrim. Gumbel PDF miring ke kanan dan menunjukkan maksimumnya pada $x = \zeta$. Probabilitas

distribusi gumbel dapat diperoleh dari fungsi distribusi kumulatif.

$$F(x) = \exp\left\{-\exp\left[-\frac{(x-\zeta)}{\beta}\right]\right\} \tag{3.58}$$

Parameter distribusi gumbel dapat diperkirakan dengan kemungkinan maksimum atau L-Momen seperti yang dijelaskan sebelumnya untuk kasus GEV yang lebih umum, tetapi cara termudah untuk menyesuaikan distribusi ini adalah dengan menggunakan metode momen. Estimasi momen untuk parameter distribusi Gumbel dihitung dengan menggunakan rata-rata sampel dan standar deviasi. persamaan estimasi adalah

$$\hat{\beta} = \frac{s\sqrt{6}}{\pi} \tag{3.59a}$$

dan

$$\hat{\zeta} = \bar{x} - \gamma\hat{\beta} \tag{3.59b}$$

dimana γ = 0.57721 adalah konstanta Euler.

Untuk k > 0, Persamaan 2.54 disebut distribusi Frechet atau Fisher-Tippett Tipe II. Distribusi ini menunjukkan apa yang disebut ekor "berat", yang berarti bahwa PDF menurun cukup lambat untuk nilai x yang besar. Konsekuensi dari ekor yang kuat adalah bahwa beberapa momen distribusi Frechet tidak terbatas. Sebagai contoh, integral yang mendefinisikan

varians (Persamaan 2.21) tidak terbatas untuk k > 1/2, dan bahkan mean [Persamaan 2.20 dengan g(x) = x] tidak terbatas untuk k > 1. Konsekuensi lain dari ekor yang kuat adalah bahwa kuantil yang terkait dengan probabilitas kumulatif yang besar (yaitu Persamaan 2.56 dengan p ≈ 1) akan cukup besar.

Khusus kasus ketiga dari distribusi GEV terjadi untuk k < 0 dan dikenal sebagai distribusi Weibull atau Fisher-Tippett tipe III. Biasanya distribusi Weibull ditulis dengan parameter shift $\zeta = 0$ dan transformasi parameter, menghasilkan PDF.

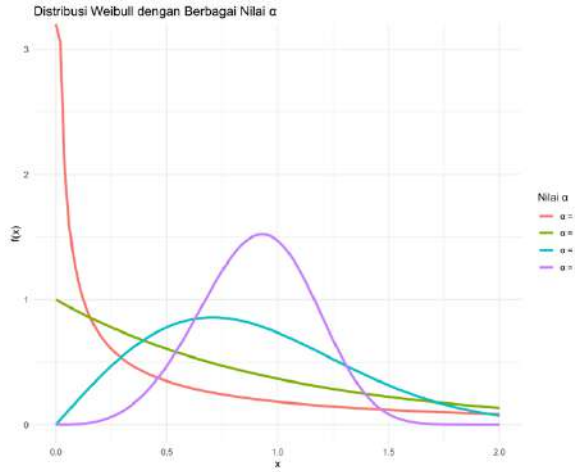
$$f(x) = \left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right], \quad x, \alpha, \beta > 0 \qquad (3.60)$$

Seperti distribusi gamma, dua parameter a dan β disebut parameter bentuk dan skala. Bentuk distribusi Weibull sama-sama dikontrol oleh dua parameter. Respon bentuk distribusi terhadap nilai a yang berbeda ditunjukkan pada **Gambar 3.12**. Seperti distribusi gamma, $a \leq 1$ menghasilkan bentuk "J" terbalik dan kemiringan positif yang kuat, dan untuk $a = 1$, distribusi Weibull juga direduksi menjadi distribusi eksponensial (Persamaan 2.45) sebagai kasus khusus. Juga bersama dengan distribusi gamma, parameter penskalaan bertindak serupa untuk meregangkan atau memampatkan bentuk dasar sepanjang sumbu x untuk nilai tertentu a. Untuk $a = 3.6$, distribusi Weibull sangat mirip dengan distribusi Gaussian.

PDF untuk distribusi Weibull dapat diintegrasikan secara analitik, menghasilkan CDF

$$f(x) = \Pr(X \leq x) = 1 - \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right] \tag{3.61}$$

Fungsi ini dapat dengan mudah dipecahkan untuk x untuk mendapatkan fungsi kuantil. Seperti halnya GEV pada umumnya, momen distribusi Weibull termasuk gamma



Gambar 3.12 Fungsi kepadatan probabilitas distribusi Weibull untuk empat nilai parameter bentuk α.

fungsi (lihat Tabel 2.6) sehingga penyesuaian parameter dengan metode momen tidak membawa keuntungan komputasi. Biasanya, distribusi Weibull dipasang menggunakan kemungkinan maksimum (lihat Bagian 4.6) atau momen-L (Stedinger et al. 1993).

Motivasi utama untuk mempelajari dan memodelkan statistik ekstrem adalah untuk memperkirakan probabilitas tahunan kejadian langka dan berpotensi berbahaya, seperti Dalam aplikasi seperti ini, asumsi teori nilai ekstrim klasik, yaitu bahwa kejadian yang mendasarinya independen dan berasal dari distribusi yang sama dan bahwa jumlah nilai individu (biasanya harian) m cukup untuk konvergensi ke GEV, tidak dapat bertemu. Yang paling bermasalah untuk penerapan teori nilai ekstrim adalah bahwa data yang mendasari seringkali tidak berasal dari distribusi yang sama, misalnya karena siklus tahunan dalam statistik nilai m (biasanya 365) dan/atau karena yang terbesar dari nilai m disebabkan oleh proses yang berbeda dalam dihasilkan di blok (tahun) yang berbeda. Misalnya, beberapa nilai curah hujan harian terbesar dapat terjadi karena pendaratan badai, beberapa mungkin terjadi karena kompleks badai petir yang besar dan bergerak lambat, dan beberapa mungkin terjadi sebagai akibat dari batas frontal yang hampir stasioner; dan statistik (yaitu PDF dasar yang sesuai dengan ini) dari berbagai proses fisik mungkin berbeda (misalnya Walshaw 2000). Pertimbangan ini tidak membatalkan GEV (Persamaan 2.54) sebagai kandidat distribusi untuk menggambarkan statistik ekstrim, dan secara empiris distribusi ini seringkali menjadi pilihan yang sangat baik bahkan ketika asumsi teori nilai ekstrim tidak terpenuhi. Namun, dalam banyak pengaturan praktis dimana asumsi klasik tidak valid, GEV tidak dijamin menjadi distribusi yang paling tepat untuk mewakili sekumpulan data nilai ekstrim. Kesesuaian GEV harus dinilai bersama dengan

kandidat distribusi lainnya untuk kumpulan data tertentu (Madsen et al. 1997; Wilks 1993). mungkin menggunakan pendekatan yang disajikan di bagian 4.6 atau 5.2.6.

Masalah praktis lain yang muncul saat bekerja dengan statistik ekstrim adalah pemilihan data ekstrim yang digunakan untuk menyesuaikan distribusi. Seperti disebutkan sebelumnya, pilihan umum adalah memilih nilai harian tunggal terbesar dalam setiap n tahun, yang dikenal sebagai rangkaian maksimum blok atau maksimum tahunan. Kemungkinan kerugian dari pendekatan ini adalah bahwa sebagian besar data tidak digunakan, termasuk nilai yang tidak terbesar pada tahun kemunculannya tetapi mungkin lebih besar dari nilai maksimum pada tahun-tahun lainnya. Pendekatan alternatif untuk menyusun sekumpulan data nilai ekstrem adalah dengan memilih nilai n terbesar terlepas dari tahun kemunculannya. Hasilnya disebut data durasi parsial dalam hidrologi. Pendekatan ini lebih dikenal sebagai peaks-over-threshold atau POT karena semua nilai yang lebih besar dari level minimum dipilih dan kami tidak dibatasi untuk memilih jumlah nilai ekstrim yang sama karena mungkin ada tahun dalam catatan klimatologis . Karena data dasar dapat menunjukkan korelasi serial yang signifikan, diperlukan kehati-hatian untuk memastikan bahwa data sub-durasi yang dipilih mewakili peristiwa yang berbeda. Secara khusus, biasanya hanya nilai berturut-turut terbesar di atas ambang pemilihan yang dimasukkan dalam kumpulan data nilai ekstrim.

Apakah tanggal durasi puncak atau parsial tahunan lebih masuk akal dalam aplikasi tertentu? Minat biasanya difokuskan pada ujung paling kanan dari distribusi nilai ekstrem yang sesuai dengan data yang sama, apakah dipilih sebagai maksimum tahunan atau puncak di atas ambang batas. Karena data durasi parsial terbesar juga akan menjadi nilai individu terbesar pada tahun-tahun kemunculannya. Biasanya, pilihan antara data durasi tahunan dan parsial paling baik dilakukan secara empiris, setelah itu distribusi nilai ekstrem yang dipasang dapat memperkirakan probabilitas nilai ekstrem dengan lebih baik (Madsen et al. 1997; Wilks 1993).

Hasil dari analisis nilai ekstrem sering kali hanyalah ringkasan kuantil yang sesuai dengan probabilitas kumulatif yang besar, seperti peristiwa dengan probabilitas terlampaui tahunan sebesar 0,01. Kecuali n cukup besar, estimasi langsung dari kuantil ekstrem ini tidak mungkin dilakukan (lihat Persamaan 1.2), dan distribusi nilai ekstrem yang pas memberikan cara yang masuk akal dan objektif untuk mengekstrapolasi probabilitas yang bisa jauh lebih besar dari 1 - 1 /n . Seringkali probabilitas ekstrim ini dinyatakan dalam rata-rata periode yang berulang.

$$R(X) = \frac{1}{\omega[1 - F(X)]} \tag{3.62}$$

Periode ulang R(x), terkait dengan kuantil x, biasanya diinterpretasikan sebagai waktu rata-rata antara kemunculan

peristiwa sebesar itu atau lebih besar. Periode ulang adalah fungsi CDF yang dievaluasi pada x dan frekuensi sampling rata-rata w. Untuk data maksimum tahunan $\omega = 1yr^{-1}$, sehingga kejadian x, yang sesuai dengan probabilitas kumulatif F(x) 0 99, akan memiliki probabilitas 1 - F(x) harus dilampaui pada tahun tertentu. Nilai x ini akan dikaitkan dengan periode ulang 100 tahun dan akan disebut sebagai peristiwa 100 tahun. Untuk data durasi parsial, ω tidak harus1 yr^{-1}, dan beberapa penulis telah menyarankan $\omega = 1.65\, yr^{-1}$ (Madsen et al. 1997; Stedinger et al. 1993). Jika seseorang memilih, misalnya, nilai harian 2n terbesar dalam n tahun terlepas dari tahun asalnya, maka $\omega = 2.0\, yr^{-1}$. Dalam hal ini, peristiwa 100 tahun akan sesuai dengan F(x) = 0.995.

Contoh 3.9 Periode Ulang dan Probabilitas Kumulatif

Seperti yang disebutkan sebelumnya, kecocokan kemungkinan maksimum dari distribusi GEV dengan data curah hujan maksimum tahunan pada Tabel 2.7 menghasilkan perkiraan parameter $\zeta = 3.50$, $\beta = 1.11$, dan $k = 0.29$. Menggunakan Persamaan 2.56 dengan probabilitas kumulatif $p = 0.5$ memberikan Median dari 3,89 inci. Ini adalah jumlah curah hujan yang memiliki peluang 50% terlampaui pada tahun tertentu. Oleh karena itu, jumlah ini terlampaui rata-rata setengah tahun dalam catatan klimatologi hipotetis yang panjang, sehingga waktu rata-rata yang memisahkan kejadian curah hujan harian sebesar ini atau lebih besar adalah dua tahun (Persamaan 2.62).

Karena data ini memiliki n = 20 tahun, median dapat langsung diperkirakan sebagai median sampel. Namun, pertimbangkan untuk memperkirakan peristiwa curah hujan 1 hari 100 tahun menggunakan data ini. Menurut Persamaan 2.62, ini sesuai dengan probabilitas kumulatif F(x) 0,99, sedangkan probabilitas kumulatif empiris yang sesuai dengan jumlah curah hujan paling ekstrem pada Tabel 2.7 dapat diperkirakan menggunakan posisi plot Tukey sebagai $p \approx 0{,}967$ (lihat Tabel 1.2). Namun, dengan menggunakan fungsi kuantil GEV Persamaan 2.56 bersama dengan Persamaan 2.62, perkiraan yang masuk akal untuk kuantitas 100 tahun dihitung pada 6,32 inci. (Jumlah presipitasi 2 dan 100 tahun yang sesuai diturunkan dari estimasi parameter momen-L, $\zeta = 3.49$, $\beta = 1.18$ dan $k = 0.32$, berturut-turut masing-masing 3,90" dan 6,33".)

Perlu ditekankan bahwa peristiwa tahun-T sama sekali tidak dijamin terjadi dalam periode tahun-T tertentu. Probabilitas kejadian tahun T terjadi pada tahun tertentu adalah 1/T , misalnya 1/T = 0.01 untuk kejadian tahun T 100. Pada tahun tertentu, terjadinya peristiwa tahun-T adalah percobaan Bernoulli dengan p = 1/T . Oleh karena itu, distribusi geometris (Persamaan 2.5) dapat digunakan untuk menghitung probabilitas menunggu beberapa tahun tertentu untuk suatu peristiwa. Interpretasi lain dari periode ulang adalah rata-rata distribusi geometris untuk latensi. Probabilitas peristiwa 100 tahun yang terjadi dalam abad yang dipilih secara acak dapat dihitung sebagai $\Pr\{X \leq 100\} = 0.634$ menggunakan Persamaan 2.5. Artinya, ada lebih dari

1/3 kemungkinan bahwa peristiwa 100 tahun tidak akan terjadi dalam 100 tahun tertentu. Demikian pula, peluang peristiwa 100 tahun tidak akan terjadi dalam 200 tahun adalah kira-kira 0,134.

3.4.6 Distribusi campuran

Distribusi parametrik yang disajikan sejauh ini dalam bab ini mungkin tidak cukup untuk data yang berasal dari lebih dari satu proses generasi atau mekanisme fisik. Contohnya adalah data suhu untuk Guayaquil pada Tabel A.3, yang histogramnya ditunjukkan pada **Gambar 2.6**. Data ini jelas bimodal; dimana punuk yang lebih kecil dan lebih hangat dalam distribusi dikaitkan dengan tahun-tahun El Niño dan punuk yang lebih besar dan lebih dingin sebagian besar terdiri dari tahun-tahun non-El Niño. Meskipun Teorema Limit Pusat menunjukkan bahwa distribusi Gaussian harus menjadi model yang baik untuk suhu rata-rata bulanan, perbedaan mencolok dalam iklim suhu Juni Guayaquil yang terkait dengan El Nino membuat distribusi Gaussian pilihan yang buruk untuk memperkirakan data ini untuk ditampilkan secara keseluruhan. Namun, distribusi Gaussian terpisah untuk tahun El Niño dan tahun non-El Niño dapat memberikan model probabilitas yang baik untuk data ini.

Kasus seperti ini adalah kandidat alami untuk representasi menggunakan distribusi campuran atau rata-rata tertimbang dari dua PDF atau lebih. Sejumlah PDF dapat digabungkan untuk membentuk distribusi campuran (Everitt dan Hand

1981; McLachlan dan Peel 2000; Titterington et al. 1985), tetapi sejauh ini distribusi campuran yang paling umum digunakan adalah rata-rata tertimbang dari PDF dua komponen.

$$f(x) = wf_1(x) + (1 + w)f_2(x) \tag{3.63}$$

PDF komponen f1(x) dan f2(x) dapat berupa distribusi apa pun, meskipun biasanya memiliki bentuk parametrik yang sama. Parameter bobot w, 0 <w< 1, menentukan kontribusi setiap kerapatan komponen terhadap campuran PDF, f(x), dan dapat diartikan sebagai probabilitas realisasi variabel acak X akan berasal dari $f_1(x)$.

Tentu saja, sifat distribusi campuran bergantung pada distribusi komponen dan parameter berat. Rata-rata hanyalah rata-rata tertimbang dari keduanya sarana komponen.

$$\mu = w\mu_1 + (1 - w)\mu_2 \tag{3.64}$$

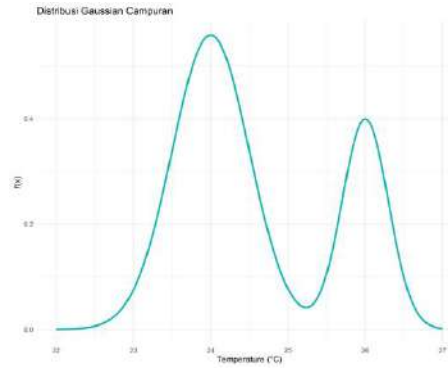
Di sisi lain, varians

$$\begin{aligned}\sigma^2 &= [w\sigma_1^2 + (1 - w)\sigma_2^2] \\ &\quad + [w(\mu_1 - \mu)^2 + (1 - w)(\mu_2 - \mu)^2] \\ &= w\sigma_1^2 + (1 - w)\sigma_2^2 + w(1 - w)(\mu_1 - \mu_2)^2\end{aligned} \tag{3.65}$$

berisi kontribusi dari varian tertimbang dari dua distribusi (suku pertama dalam tanda kurung siku di baris pertama) ditambah sebaran tambahan yang dihasilkan dari selisih dua

mean (suku kedua dalam tanda kurung siku). Distribusi campuran jelas mampu merepresentasikan bimodalitas (atau, jika campuran terdiri dari tiga atau lebih distribusi komponen, multimodalitas), tetapi distribusi campuran juga dapat unimodal jika perbedaan antara rata-rata komponen cukup kecil relatif terhadap standar deviasi atau varians dari komponen.

Biasanya, distribusi campuran dipasang menggunakan metode kemungkinan maksimum menggunakan algoritma EM (lihat Bagian 4.6.3). **Gambar 3.13** menunjukkan PDF untuk kecocokan kemungkinan maksimum campuran 2 distribusi Gaussian dengan data suhu Guayaquil June pada Tabel A.3 dengan parameter μ_1 = 24.34°C, σ_1= 0.46°C, μ_2 = 26.48°C, σ_2 = 0.26°C dan w = 0.80 (lihat Contoh 4.13). Di sini μ_1 dan σ_1 adalah parameter yang pertama (lebih dingin dan banyak lagi kemungkinan) Distribusi Gaussian, $f_1(x)$, dan μ_2 dan σ_2 adalah parameter distribusi Gaussian kedua (lebih hangat dan kurang mungkin), $f_2(x)$. Perpaduan PDF pada **Gambar 3.13**



Gambar 3.13 Fungsi kerapatan peluang untuk campuran (Persamaan 2.63) dari dua distribusi Gaussian cocok dengan data suhu Guayaquil bulan Juni (Tabel A3). Hasilnya sangat mirip dengan perkiraan densitas kernel yang diperoleh dari data yang sama, Gambar 1.8b

muncul sebagai tambahan sederhana (berbobot) dari dua komponen Gaussian, mirip dengan konstruksi perkiraan kepadatan kernel untuk data yang sama pada Gambar 1.8, sebagai jumlah dari kernel yang diskalakan, yang merupakan fungsi kepadatan probabilitas. Faktanya, campuran Gaussian pada **Gambar 3.13** serupa dengan perkiraan densitas kernel yang diturunkan dari data yang sama pada **Gambar 2.8**b. Sarana dari distribusi Gaussian dua komponen dipisahkan dengan baik relatif terhadap pencar yang dicirikan oleh dua standar deviasi, menghasilkan distribusi campuran menjadi bimodal yang kuat.

Distribusi Gaussian adalah pilihan paling umum untuk komponen distribusi campuran, tetapi campuran distribusi eksponensial (Persamaan 2.45) juga penting dan umum

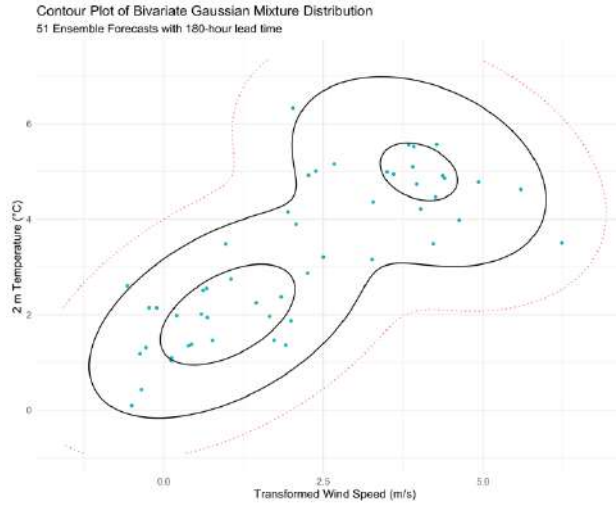
digunakan. Secara khusus, distribusi campuran dari dua distribusi eksponensial disebut distribusi eksponensial campuran dengan PDF.

$$f(x) = \frac{w}{\beta_1}\exp\left(-\frac{x}{\beta_1}\right) + \frac{1-w}{\beta_2}\exp\left(-\frac{x}{\beta_2}\right) \qquad (3.66)$$

Distribusi eksponensial campuran telah terbukti sangat sesuai untuk data curah hujan harian bukan nol (Woolhiser dan Roldan 1982; Foufoula-Georgiou dan Lettenmaier 1987; Wilks 1999a) dan sangat berguna untuk simulasi (lihat Bagian 4.7) curah hujan harian berkorelasi spasial jumlah (Wilks 1998).

Distribusi campuran tidak terbatas pada kombinasi PDF kontinu univariat. Bentuk Persamaan 2.63 juga dapat dengan mudah digunakan untuk membentuk campuran fungsi distribusi probabilitas diskrit atau campuran distribusi gabungan multivariat. Sebagai contoh, **Gambar 3.14** menunjukkan campuran dari dua distribusi Gaussian bivariat (Persamaan 2.33) yang sesuai dengan perkiraan ansambel 51 anggota (lihat Bagian 6.6) untuk suhu dan kecepatan angin. Distribusi dipasang menggunakan algoritme kemungkinan maksimum untuk campuran Gaussian multivariat yang dijelaskan dalam Smyth et al. (1999) dan Hannachi dan O'Neill (2001). Meskipun distribusi campuran multivariat cukup fleksibel dalam mengakomodasi data yang terlihat tidak biasa, fleksibilitas ini datang dengan harga yang membutuhkan sejumlah besar parameter untuk diestimasi,

sehingga penggunaan model probabilitas yang relatif canggih dari jenis ini dapat dibatasi oleh ukuran sampel yang tersedia. . Distribusi campuran pada **Gambar 3.14** memerlukan 11 parameter untuk mencirikannya: dua rata-rata, dua varian, dan korelasi untuk masing-masing dari dua distribusi komponen bivariat, ditambah parameter pembobot w.



Gambar 3.14 Plot kontur PDF dari distribusi campuran Gaussian bivariat yang dipasang pada ansambel 51 prakiraan untuk suhu 2 m dan kecepatan angin 10 m, dibuat dengan waktu tunggu 180 jam. Kecepatan angin pertama kali ditransformasikan akar kuadrat untuk membuat distribusi univariatnya lebih Gaussian. Titik menandai prakiraan individu dari 51 anggota ansambel. Dua densitas Gaussian bivariat konstituen $f_1(x)$ dan $f_2(x)$ masing-masing berpusat pada 1 dan 2, dan garis halus menunjukkan kurva level dari campurannya $f(x)$ yang dibentuk dengan w = 0.57. Interval kontur tetap adalah 0,05 , dan garis putus-putus berat dan ringan masing-masing adalah 0,01 dan 0,001. Diadaptasi dari Wilks (2002b).

3.5 Penilaian Kualitatif tentang Goodness of Fit

Setelah memasang distribusi parametrik ke tumpukan data, lebih dari sekadar kepentingan untuk memverifikasi bahwa model probabilistik teoretis memberikan deskripsi yang memadai. Menyesuaikan distribusi yang tidak tepat dapat menyebabkan kesimpulan yang salah. Metode kuantitatif untuk menaksir kedekatan distribusi pas dengan data dasar mengacu pada gagasan dari pengujian hipotesis formal, dan beberapa metode seperti itu disajikan di Bagian 5.2.5. Bagian ini menjelaskan beberapa metode kualitatif dan grafis yang berguna untuk mengidentifikasi kesesuaian secara subyektif. Prosedur-prosedur ini bersifat instruktif bahkan ketika uji kecocokan formal harus dihitung. Tes formal mungkin menunjukkan kecocokan yang tidak memadai, tetapi mungkin tidak memberi tahu analis tentang sifat spesifik dari masalah tersebut. Perbandingan grafis dari data dan distribusi yang sesuai memungkinkan diagnosis dimana dan bagaimana representasi teoretis mungkin tidak mencukupi

3.5.1 Superposisi Distribusi Parametrik Terpasang dan Histogram Data

Mungkin cara paling sederhana dan paling intuitif untuk membandingkan distribusi parametrik yang sesuai dengan data yang mendasarinya adalah dengan melapisi distribusi yang sesuai dan histogram. Dengan cara ini, penyimpangan

besar dari data mudah dikenali. Jika data cukup banyak, penyimpangan histogram akibat variasi sampel tidak akan terlalu mengganggu.

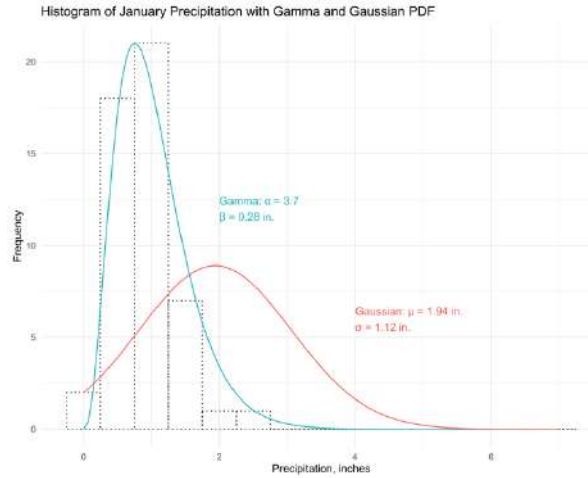
Untuk data diskrit, fungsi distribusi probabilitas sudah sangat mirip dengan histogram. Baik histogram dan fungsi distribusi probabilitas menetapkan probabilitas ke serangkaian hasil yang terpisah. Membandingkan keduanya hanya membutuhkan nilai data diskrit yang sama, atau rentang nilai data, diplot dan bahwa histogram dan fungsi distribusi diskalakan secara sebanding. Kondisi kedua ini dipenuhi dengan memplot histogram sebagai frekuensi relatif dan bukan frekuensi absolut pada sumbu vertikal. **Gambar 3.2** adalah contoh fungsi distribusi probabilitas Poisson yang dihamparkan pada histogram jumlah tornado tahunan yang diamati di negara bagian New York.

Prosedur untuk menempatkan PDF kontinu pada histogram sepenuhnya serupa. Kendala dasarnya adalah bahwa integral dari setiap fungsi kerapatan probabilitas harus merupakan kesatuan pada seluruh rentang variabel acak. Artinya, Persamaan 2.17 dipenuhi oleh semua fungsi kerapatan probabilitas. Salah satu pendekatan untuk menyesuaikan histogram dan fungsi densitas adalah mengubah skala fungsi densitas. Faktor penskalaan yang benar diperoleh dengan menghitung luas yang ditempati secara kolektif oleh semua batang dalam plot histogram. Jika kita menyatakan luas ini dengan A, mudah untuk melihat bahwa mengalikan fungsi kerapatan $f(x)$ dengan A menghasilkan kurva yang luasnya

juga A, karena A dapat dikeluarkan dari integral sebagai konstanta: A • $f(x)dx$ = A • $f(x)dx$ = A • 1 = A. Perhatikan bahwa juga memungkinkan untuk menskala ulang ketinggian histogram sehingga total area yang terdapat dalam batang adalah 1. Pendekatan terakhir ini lebih tradisional dalam statistik karena histogram dipandang sebagai perkiraan fungsi kerapatan.

Contoh 3.10 Melapisi PDF pada histogram

Gambar 3.15 mengilustrasikan prosedur untuk melapisi distribusi yang sesuai dan histogram untuk total curah hujan Januari 1933–1982 di Ithaca dari Tabel A.2. Berikut adalah data n = 50 tahun dan lebar nampan untuk histogram (konsisten dengan Persamaan 1.12) adalah 0,5 inci, sehingga luas yang ditempati oleh persegi panjang histogram adalah A = (50)(0 5) = 25. PDF dihamparkan pada histogram ini untuk distribusi gamma menggunakan Persamaan 2.41 atau 4.43a



Gambar 3.15 Histogram data curah hujan Januari 1933–1982 di Ithaca dari Tabel A2 dengan PDF gamma (padat) dan Gaussian (rusak). Masing-masing dari dua fungsi kerapatan telah dikalikan dengan A = 25 karena lebar bin adalah 0,5 inci dan terdapat 50 pengamatan. Rupanya distribusi gamma memberikan representasi data yang masuk akal. Distribusi Gaussian kurang mewakili ekor kanan dan menyiratkan probabilitas presipitasi negatif yang tidak nol.

(kurva padat) dan menyesuaikan distribusi Gaussian dengan menyesuaikan sampel dan momen distribusi (kurva putus-putus). Dalam kedua kasus, PDF (Persamaan 2.38 dan 4.23, masing-masing) telah dikalikan dengan 25 sehingga luasnya sama dengan histogram. Jelas bahwa distribusi Gaussian simetris adalah pilihan yang buruk untuk mewakili data curah hujan yang condong positif ini, karena jumlah curah hujan terbesar diberi probabilitas terlalu kecil dan jumlah curah hujan negatif yang tidak mungkin diberi probabilitas yang tidak dapat diabaikan. Distribusi gamma mewakili data ini

jauh lebih akurat dan memberikan ringkasan yang cukup masuk akal dari variasi data dari tahun ke tahun. Kecocokan tampaknya paling buruk untuk nampan 0,75"-1,25" dan 1,25"-1,75", meskipun hal ini dapat dengan mudah dihasilkan dari variasi sampel. Kumpulan data yang sama juga digunakan di Bagian 5.2.5 untuk menguji secara formal kesesuaian kedua distribusi ini.

3.5.2 Diagram Kuantil-Kuantil (Q–Q)

Plot atau diagram kuantil-kuantil (Q-Q) membandingkan empiris (data) dan CDF yang dipasang dalam hal nilai dimensi variabel (kuantil empiris). Hubungan antara pengamatan dari variabel acak x dan distribusi yang sesuai ditetapkan melalui fungsi kuantil atau kebalikan dari CDF (Persamaan 2.19), yang dievaluasi pada perkiraan tingkat probabilitas kumulatif.

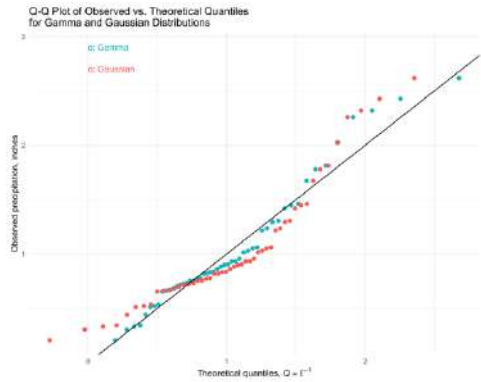
Bagan Q-Q adalah plot pencar. Setiap pasangan koordinat yang menentukan lokasi suatu titik terdiri dari nilai data dan perkiraan yang sesuai untuk nilai data tersebut, yang diturunkan dari fungsi kuantil dari distribusi yang dipasang. Mengambil rumus posisi petak Tukey (lihat Tabel 1.2) sebagai estimator untuk probabilitas kumulatif empiris (walaupun yang lain dapat digunakan), setiap titik pada plot Q-Q akan memiliki koordinat Cartesian $(F^{-1}[(i - 1/3)/(n - 1/3)], x_{(i)})$. Jadi, titik ke-i dalam diagram Q–Q ditentukan oleh nilai data terkecil ke-i $x_{(i)}$ dan nilai variabel acak yang sesuai dengan probabilitas sampling kumulatif p = (i - 1/3)/(n +1/3)

dalam distribusi. Plot Q-Q untuk distribusi ini yang mewakili data dengan sempurna akan memiliki semua titik pada garis diagonal 1:1

Gambar 3.16 menunjukkan plot Q-Q yang membandingkan kecocokan distribusi gamma dan Gaussian dengan data curah hujan Ithaca untuk Januari 1933–1982 pada Tabel A.2 (perkiraan parameter ditunjukkan pada **Gambar 3.15**). **Gambar 3.16** menunjukkan bahwa distribusi gamma yang dipasang cocok dengan data pada sebagian besar rentangnya karena fungsi kuantil yang dievaluasi menggunakan perkiraan probabilitas kumulatif empiris cukup dekat dengan nilai data yang diamati dan memberikan titik yang sangat dekat dengan garis 1:1 . Distribusi pas tampaknya meremehkan beberapa titik terbesar, menunjukkan bahwa ekor distribusi gamma pas mungkin terlalu tipis, meskipun setidaknya beberapa perbedaan ini mungkin disebabkan oleh variasi sampel.

Di sisi lain, **Gambar 3.16** menunjukkan bahwa kecocokan Gaussian dengan data ini jauh lebih rendah. Terutama, ekor kiri dari distribusi Gaussian pas terlalu kuat, sehingga kuantil teoretis terkecil terlalu kecil, dan dua yang terkecil sebenarnya negatif. Melalui sebagian besar distribusi, kuantil Gaussian lebih jauh dari garis 1: 1 daripada kuantil gamma, menunjukkan kecocokan yang kurang akurat, dan di ujung kanan, distribusi Gaussian meremehkan kuantil terbesar bahkan lebih dari distribusi gamma .

Dimungkinkan juga untuk membandingkan distribusi pas dan empiris dengan membalikkan logika plot Q-Q dan memplot sebar probabilitas kumulatif empiris (diperkirakan menggunakan posisi plot, Tabel 1.2) versus CDF yang dipasang, $F(x)$



Gambar 3.16 Plot kuantil-kuantil untuk gamma (o) dan Gaussian (x) sesuai dengan curah hujan Januari 1933–1982 di Ithaca pada Tabel A2. Jumlah curah hujan yang diamati berada pada posisi vertikal dan jumlah yang berasal dari distribusi yang dipasang menggunakan posisi plot Tukey berada pada posisi horizontal. Garis diagonal menunjukkan korespondensi 1:1.

dievaluasi dengan nilai data yang sesuai. Plot jenis ini disebut plot probabilitas-probabilitas atau P-P. P-Plot tampaknya lebih jarang digunakan daripada plot Q-Q, mungkin karena perbandingan nilai data dimensi bisa lebih intuitif daripada perbandingan probabilitas kumulatif. Plot P-P juga kurang sensitif terhadap perbedaan ujung ekstrim dari suatu distribusi, yang seringkali paling menarik. Grafik Q-Q dan P-P

termasuk dalam kelas grafik yang lebih luas yang dikenal sebagai grafik probabilitas.

3.6 Pemasangan Parameter Kemungkinan Maksimum

3.6.1 Fungsi kemungkinan (Likelihood)

Untuk banyak distribusi, pemasangan parameter menggunakan metode momen sederhana menghasilkan hasil yang lebih rendah yang dapat menyebabkan kesimpulan dan ekstrapolasi yang menyesatkan. Metode kemungkinan maksimum adalah alternatif yang serbaguna dan penting. Seperti namanya, metode ini mencoba menemukan nilai parameter distribusi yang memaksimalkan fungsi kemungkinan. Prosedur mengikuti dari gagasan bahwa probabilitas adalah ukuran sejauh mana data mendukung nilai-nilai tertentu dari parameter (misalnya Lindgren 1976). Interpretasi Bayesian dari prosedur (kecuali untuk ukuran sampel kecil) adalah bahwa penaksir kemungkinan maksimum adalah nilai yang paling mungkin untuk parameter yang diberikan data yang diamati.

Secara notasi, fungsi kemungkinan untuk pengamatan tunggal x terlihat identik dengan fungsi kerapatan probabilitas (atau distribusi probabilitas untuk variabel diskrit), dan perbedaan antara keduanya dapat membingungkan pada awalnya.

Perbedaannya adalah PDF adalah fungsi data untuk nilai tetap dari parameter, sedangkan fungsi probabilitas adalah fungsi dari parameter yang tidak diketahui untuk nilai tetap dari data (sudah diamati). Sama seperti PDF bersama dari n variabel independen adalah produk dari n PDF individu, fungsi probabilitas untuk parameter distribusi yang diberikan sampel n nilai data independen hanyalah produk dari n fungsi probabilitas individu. Misalnya, fungsi probabilitas untuk parameter Gaussian μ dan sampel n pengamatan yang diberikan adalah $x_i, i = 1, \dots, n$.

$$\Lambda(\mu, \sigma) = \sigma^{-n}\left(\sqrt{2\pi}\right)^{-n} \prod_{i=1}^{n} \exp\left[-\frac{(x_1 - \mu)^2}{2\sigma^2}\right] \qquad (3.67)$$

Di sini, pi kapital menunjukkan perkalian suku-suku bentuk yang diberikan di sebelah kanannya, analog dengan penambahan yang ditunjukkan oleh notasi sigma kapital. Bahkan, probabilitas dapat berupa fungsi apa pun yang sebanding dengan Persamaan 2.67, sehingga faktor konstanta yang melibatkan akar kuadrat dari 2π dapat dihilangkan karena tidak bergantung pada salah satu parameter. Itu dimasukkan untuk menekankan hubungan antara Persamaan 2.67 dan 4.23. Sisi kanan Persamaan 2.67 terlihat sama dengan PDF bersama untuk n variabel Gaussian independen, kecuali bahwa parameter μ dan a adalah variabel dan x_i menunjukkan konstanta tetap. Secara geometris, Persamaan 2.67 menggambarkan area di atas μ, sebuah bidang yang mengambil nilai maksimum pada pasangan nilai parameter

tertentu tergantung pada kumpulan data tertentu yang diberikan oleh nilai x_i.

Biasanya lebih nyaman untuk bekerja dengan logaritma dari fungsi kemungkinan, yang disebut kemungkinan log. Karena logaritma adalah fungsi yang meningkat secara ketat, nilai parameter yang sama memaksimalkan fungsi kemungkinan dan kemungkinan log. Fungsi kemungkinan log untuk parameter Gaussian sesuai dengan Persamaan 2.67

$$L(\mu, \sigma) = \ln[\Lambda(\mu, \sigma)]$$

$$= -n \ln(\sigma) - n \ln\left(\sqrt{2\pi}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \qquad (3.68)$$

dimana lagi suku dengan 2π tidak mutlak diperlukan untuk menemukan maksimum fungsi karena tidak bergantung pada parameter μ dan a. Setidaknya secara konseptual, memaksimalkan kemungkinan log adalah latihan komputasi sederhana. Untuk distribusi Gaussian, latihannya sangat mudah karena maksimisasi dapat dilakukan secara analitis. Derivasi Persamaan 2.68 sehubungan dengan parameter yang diberikan.

$$\frac{\partial L(\mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2} \left[\sum_{i=1}^{n} x_i - n\mu \right] \qquad (3.69a)$$

dan

$$\frac{\partial L(\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{3.69b}$$

Tetapkan masing-masing turunan ini sama dengan nol dan diperoleh hasilnya.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.70a}$$

dan

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2} \tag{3.70b}$$

Ini adalah estimator kemungkinan maksimum (MLE) untuk distribusi Gaussian, yang mudah dikenali sangat mirip dengan estimator momen. Satu-satunya perbedaan adalah pembagi pada Persamaan 2.70b, yaitu n dan bukan n-1. Pembagi n-1 sering digunakan saat menghitung standar deviasi sampel karena pilihan ini memberikan perkiraan nilai populasi yang tidak bias. Perbedaan ini menunjukkan fakta bahwa estimator kemungkinan maksimum untuk distribusi tertentu mungkin tidak bias. Dalam hal ini, estimasi standar deviasi (Persamaan 2.70b) akan cenderung terlalu kecil secara rata-rata, karena xi, secara rata-rata, lebih dekat dengan rata-rata sampel yang

dihitung dari Persamaan 2.70a daripada rata-rata sebenarnya, meskipun perbedaan ini kecil untuk n besar.

3.6.2 Metode Newton-Raphson

MLE untuk distribusi Gaussian agak tidak biasa karena dapat dihitung secara analitik. Lebih khas bahwa perkiraan ke MLE dihitung secara iteratif. Pendekatan umum adalah memikirkan memaksimalkan kemungkinan log sebagai masalah pencarian akar nonlinier yang perlu diselesaikan dengan menggunakan generalisasi multidimensi dari metode Newton-Raphson (mis. Press et al. 1986). Pendekatan ini mengikuti ekspansi Taylor yang terpotong dari turunan fungsi log-likelihood.

$$L'(\theta^*) \approx L'(\theta) + (\theta^* - \theta)L''(\theta) \tag{3.71}$$

dimana θ menunjukkan vektor umum dari parameter distribusi dan θ^* adalah nilai sebenarnya yang akan didekati. Karena merupakan turunan dari fungsi probabilitas log-likelihood $L`(\theta^*)$ yang harus dicari akarnya, Persamaan 2.71 memerlukan perhitungan turunan kedua dari probabilitas log-likelihood $L``(\theta)$. Menyetel Persamaan 2.71 sama dengan nol (untuk menemukan maksimum dalam probabilitas log-likelihood L) dan menata ulang memberikan ekspresi yang menjelaskan algoritme untuk prosedur berulang,

$$\theta^* = \theta - \frac{L'(\theta)}{L''(\theta)} \tag{3.72}$$

Dimulai dengan tebakan awal θ, kami menghitung set tebakan yang diperbarui, θ^*, yang pada gilirannya digunakan sebagai tebakan untuk iterasi berikutnya.

Contoh 3.11 Algoritma untuk Estimasi Kemungkinan Maksimum dari Parameter Distribusi Gamma

Dalam prakteknya, menggunakan Persamaan 2.72 agak rumit dengan fakta bahwa biasanya lebih dari satu parameter perlu diestimasi, jadi $L`(\theta)$ adalah vektor turunan pertama dan $L``(\theta)$ adalah matriks turunan kedua. Sebagai ilustrasi, pertimbangkan distribusi gamma (Persamaan 2.38). Untuk distribusi ini, Persamaan 2.72.

$$\begin{bmatrix}\alpha^*\\ \beta^*\end{bmatrix} = \begin{bmatrix}\alpha\\ \beta\end{bmatrix} - \begin{bmatrix}\dfrac{\partial^2 L}{\partial \alpha^2} & \dfrac{\partial^2 L}{\partial \alpha \partial \beta}\\ \dfrac{\partial^2 L}{\partial \beta \partial \alpha} & \dfrac{\partial^2 L}{\partial \beta^2}\end{bmatrix}^{-1} \begin{bmatrix}\dfrac{\partial L}{\partial \alpha}\\ \dfrac{\partial L}{\partial \beta}\end{bmatrix}$$

$$= \begin{bmatrix}\alpha\\ \beta\end{bmatrix} - \begin{bmatrix}-n\Gamma''(\alpha) & -\dfrac{n}{\beta}\\ -\dfrac{n}{\beta} & \dfrac{n\alpha}{\beta^2} - \dfrac{2\sum x}{\beta^3}\end{bmatrix}^{-1}$$

$$\times \begin{bmatrix}\ln \sum(x) - n\ln(\beta) - n\,\Gamma'(\alpha)\\ \dfrac{\sum x}{\beta^2} - \dfrac{n\alpha}{\beta}\end{bmatrix} \qquad (3.73)$$

dimana $\Gamma`(a)$ dan $\Gamma``(a)$ adalah turunan pertama dan kedua dari fungsi gamma (Persamaan 2.7) yang perlu dievaluasi atau didekati secara numerik (mis. Abramowitz dan Stegun 1984). Notasi aljabar matriks pada persamaan ini dijelaskan pada Bab 9. Persamaan 2.73 akan diimplementasikan dimulai dengan penaksiran awal untuk parameter a dan β, mungkin dengan menggunakan penaksir momen (Persamaan 2.39). Nilai yang diperbarui, a^* dan β^* akan dihasilkan dari penerapan Persamaan 2.73. Nilai yang diperbarui kemudian akan disubstitusi ke ruas kanan Persamaan 2.73, dan prosesnya diulang sampai algoritma konvergen. Konvergensi dapat didiagnosis dengan perkiraan parameter yang berubah cukup sedikit di antara iterasi, mungkin dengan sebagian kecil dari satu persen. Perhatikan bahwa dalam praktiknya algoritma Newton-Raphson dapat melebihi probabilitas maksimum pada iterasi tertentu, yang dapat menyebabkan penurunan dari satu iterasi ke iterasi berikutnya dalam perkiraan saat ini ke kemungkinan log. Seringkali algoritma Newton-Raphson diprogram untuk memeriksa penurunan probabilitas tersebut dan untuk mencoba perubahan yang lebih kecil pada parameter yang diestimasi (walaupun dalam arah yang sama diberikan dalam kasus ini oleh Persamaan 2.73).

3.6.3 Algoritma EM

Estimasi kemungkinan maksimum menggunakan metode Newton-Raphson umumnya cepat dan efektif dalam aplikasi dimana relatif sedikit parameter yang perlu diestimasi.

Namun, untuk masalah dengan lebih dari tiga parameter, perhitungan yang diperlukan dapat berkembang secara dramatis. Lebih buruk lagi, iterasi bisa sangat tidak stabil (terkadang menghasilkan parameter yang diperbarui "liar" θ^* jauh dari probabilitas maksimum nilai yang dicari), kecuali tebakan awal sangat dekat dengan nilai yang benar Prosedur estimasi itu sendiri hampir tidak diperlukan.

Sebuah alternatif untuk Newton-Raphson yang tidak memiliki masalah ini adalah algoritma EM atau Maksimalisasi Ekspektasi (McLachlan dan Krishnan 1997). Nyatanya, agak tidak tepat menyebut algoritme EM sebagai "algoritma" karena tidak ada spesifikasi eksplisit (seperti Persamaan 2.72 untuk metode Newton-Raphson) dari langkah-langkah yang diperlukan untuk mengimplementasikannya secara umum. Sebaliknya, ini lebih merupakan pendekatan konseptual yang perlu disesuaikan dengan masalah tertentu.

Algoritma EM diformulasikan sebagai bagian dari estimasi parameter untuk data "tidak lengkap". Oleh karena itu, pada satu tingkat ini sangat cocok untuk situasi dimana beberapa data mungkin hilang atau tidak teramati di atas atau di bawah ambang batas yang diketahui (data tersensor dan data terpotong), atau terekam secara tidak tepat karena binning kasar. Situasi seperti itu dengan mudah ditangani oleh algoritme EM jika masalah estimasi sederhana (misalnya direduksi menjadi solusi analitik seperti Persamaan 2.70) saat datanya "lengkap". Secara lebih umum, masalah estimasi biasa (yaitu tidak inheren "tidak lengkap") dapat ditangani

dengan algoritma EM jika adanya beberapa data tambahan yang tidak diketahui (dan mungkin hipotetis atau tidak dapat dikenali) memungkinkan maksimum sederhana (misalnya analitis) untuk dirumuskan. metode estimasi. Seperti metode Newton-Raphson, algoritme EM membutuhkan perhitungan berulang dan oleh karena itu estimasi awal dari parameter yang akan diestimasi. Jika algoritma EM dapat diformulasikan untuk masalah estimasi kemungkinan maksimum, kesulitan yang dihadapi oleh pendekatan Newton-Raphson tidak muncul, dan khususnya kemungkinan log yang diperbarui tidak akan berkurang dari iterasi ke iterasi, terlepas dari berapa banyak parameter yang diestimasi. sekaligus. Misalnya, konstruksi **Gambar 3.14** membutuhkan estimasi simultan dari 11 parameter, yang akan menjadi tidak praktis secara numerik dengan menggunakan pendekatan Newton-Raphson kecuali jawaban yang benar pada awalnya diketahui dengan pendekatan yang baik.

Persis apa yang merupakan jenis data "lengkap" yang memungkinkan kelancaran penggunaan mesin algoritma EM akan bervariasi dari satu masalah ke masalah lainnya dan mungkin memerlukan beberapa kreativitas untuk mendefinisikannya. Oleh karena itu, meskipun contoh berikut mengilustrasikan sifat proses, tidak praktis untuk menguraikan proses di sini secara umum cukup untuk dijadikan sebagai panduan mandiri untuk penggunaannya. Contoh lain penggunaannya dalam literatur sains atmosfer termasuk Hannachi dan O'Neill (2001), Katz dan Zheng (1999), Sansom dan Thomson (1992), dan Smyth et al.

(1999). Karya sumber aslinya adalah Dempster et al. (1977), dan pengobatan sepanjang buku yang otoritatif adalah McLachlan dan Krishnan (1997).

Contoh 3.12 Menyesuaikan Campuran Dua Distribusi Gaussian dengan Algoritma EM

Gambar 3.13 menunjukkan kesesuaian PDF dengan data suhu Guayaquil pada Tabel A.3 dengan asumsi distribusi campuran dalam bentuk Persamaan 2.63, dengan kedua komponen PDF $f_1(x)$ dan $f_2(x)$ diasumsikan sebagai Gaussian (Persamaan 2.23). Seperti disebutkan sehubungan dengan **Gambar 3.13**, metode pemasangan kemungkinan maksimum menggunakan algoritma EM.

Salah satu interpretasi Persamaan 2.63 adalah bahwa setiap datum x diambil dari $f_1(x)$ atau $f_2(x)$, dengan total frekuensi relatif w dan $(1 - w)$, berturut-turut. Tidak diketahui x mana yang ditarik dari PDF yang mana, tetapi jika informasi yang lebih lengkap ini tersedia, maka menyesuaikan campuran dua distribusi Gaussian yang diberikan dalam Persamaan 2.63 akan sederhana: menentukan parameter μ_1 dan σ_1 PDF $f_1(x)$ dapat diestimasi hanya berdasarkan data $f_1(x)$ menggunakan Persamaan 2.70, parameter μ_2 dan σ_2 yang mendefinisikan PDF $f_2(x)$ dapat diestimasi menggunakan Persamaan 2.70 hanya data yang diestimasi dari $f_2(x)$, dan parameter pencampuran w dapat diperkirakan sebagai sebagian kecil dari data $f_1(x)$.

Sekalipun label yang mengidentifikasi x tertentu yang diambil dari $f_1(x)$ atau $f_2(x)$ tidak tersedia (sehingga kumpulan datanya "tidak lengkap"), estimasi parameter dapat dilanjutkan dengan menggunakan nilai yang diharapkan dari pengidentifikasi hipotetis ini di masing-masing langkah iterasi. Jika variabel pengidentifikasi hipotetis adalah biner (sama dengan 1 untuk $f_1(x)$ dan sama dengan 0 untuk $f_2(x)$), diberi nilai data x_i, nilai yang diharapkan akan sesuai dengan probabilitas bahwa x_i diambil dari $f_1(x)$. Parameter campuran w akan sama dengan hipotetis rata-rata dari n variabel biner.

Persamaan 11.17 menentukan nilai yang diharapkan dari variabel indikator hipotetis (yaitu n probabilitas bersyarat) dalam bentuk dua PDFS $f_1(x)$ dan $f_2(x)$ dan parameter pencampuran w:

$$P(f_1|x_i) = \frac{wf_1(x_i)}{wf_1(x_i) + (1+w)f_2(x_i)}, i = 1, \dots, n \qquad (3.74)$$

Setelah probabilitas n dihitung, selanjutnya adalah perkiraan yang diperbarui untuk parameter campuran

$$w = \frac{1}{n}\sum_{i=1}^{n} P(f_1|x_i) \qquad (3.75)$$

Persamaan 2.74 dan 4.75 mendefinisikan bagian E (atau ekspektasi) dari penerapan algoritme EM ini, dimana ekspektasi statistik dihitung untuk data keanggotaan grup biner yang tidak diketahui (dan hipotetis). Setelah

probabilitas ini dihitung, bagian -M (atau -maksimalkan) dari algoritme EM adalah kemungkinan maksimum biasa Estimasi (Persamaan 2.70, untuk pemasangan distribusi Gaussian) menggunakan jumlah yang diharapkan ini alih-alih pasangan "data lengkap" yang tidak diketahui:

$$\hat{\mu}_1 = \frac{1}{nw}\sum_{i=1}^{n} P(f_1|x_i)x_i \tag{3.76a}$$

$$\hat{\mu}_2 = \frac{1}{n(1-w)}\sum_{i=1}^{n} [1 - P(f_1|x_i)]x_i \tag{3.76b}$$

$$\hat{\mu}_2 = \left[\frac{1}{nw}\sum_{i=1}^{n} P(f_1|x_i)(x_i - \hat{\mu}_1)^2\right]^{\frac{1}{2}} \tag{3.76c}$$

dan

$$\hat{\sigma}_2 = \left[\frac{1}{n(1-w)}\sum_{i=1}^{n} [1 - P(f_1|x_i)](x_i - \hat{\mu}_2)^2\right]^{\frac{1}{2}} \tag{3.76d}$$

Artinya, Persamaan 2.76 mengimplementasikan Persamaan 2.70 untuk masing-masing dari dua distribusi Gaussian $f_1(x)$ dan $f_2(x)$, menggunakan nilai yang diharapkan untuk variabel indikator hipotetis, daripada mengurutkan x menjadi dua kelompok yang terpisah. Jika label hipotetis ini diketahui, pengurutan ini akan sesuai dengan nilai $P(f_1|x_i)$ yang sama dengan indikator biner yang sesuai, sehingga Persamaan 2.75

akan menjadi frekuensi relatif pengamatan $f_1(x)$ dan setiap x_i akan berkontribusi hanya pada Persamaan 2.76a dan 4.76c atau pada Persamaan 2.76b dan 4.76d.

Implementasi algoritma EM untuk pendugaan parameter campuran PDF untuk dua distribusi Gaussian pada Persamaan 2.63 dimulai dengan pendugaan awal untuk kelima parameter distribusi μ_1, σ_1, μ_2 dan σ_2 dan w. Pendugaan awal ini digunakan pada Persamaan 2.74 dan 4.75 digunakan untuk mendapatkan perkiraan awal untuk probabilitas selanjutnya $P(f_1|x_i)$ dan parameter pencampuran yang diperbarui w. Nilai yang diperbarui untuk dua rata-rata dan dua standar deviasi kemudian diperoleh dengan menggunakan Persamaan 2.76 dan proses tersebut diulang hingga konvergensi. Tebakan awal tidak perlu terlalu bagus. Sebagai contoh, Tabel 2.8 menguraikan kemajuan algoritma EM dalam menyesuaikan distribusi campuran yang ditunjukkan pada **Gambar 3.13**, dimulai dengan perkiraan awal yang agak buruk $\mu_1 = 22$°C, $\mu_2 = 28$°C, dan $\sigma_1 = \sigma_2 = 1$°C dan $w = 0.5$. Catat bahwa perkiraan awal untuk kedua rata-rata bahkan tidak berada dalam jangkauan data.

Tabel 3.8 Kemajuan algoritme EM selama tujuh iterasi yang diperlukan agar sesuai dengan campuran PDF Gaussian yang ditunjukkan pada **Gambar 3.13**

Iteration	w	μ_1	μ_2	α_1	α_2	Log-likelihood
0	0.50	22.00	28.00	1.00	1.00	−79.73
1	0.71	24.26	25.99	0.42	0.76	−22.95
2	0.73	24.28	26.09	0.43	0.72	−22.72

3	0.75	24.30	26.19	0.44	0.65	−22.42
4	0.77	24.31	26.30	0.44	0.54	−21.92
5	0.79	24.33	26.40	0.45	0.39	−21.09
6	0.80	24.34	26.47	0.46	0.27	−20.49
7	0.80	24.34	26.48	0.46	0.26	−20.48

Tabel 2.8 menunjukkan bahwa rata-rata yang diperbarui cukup dekat dengan nilai akhirnya setelah hanya satu iterasi, dan algoritme telah konvergen setelah tujuh iterasi. Kolom terakhir dalam tabel ini menunjukkan bahwa kemungkinan log meningkat secara monoton dengan setiap iterasi

3.6.4 Distribusi Sampling Estimasi Kemungkinan Maksimum

Meskipun perkiraan kemungkinan maksimum membutuhkan perhitungan yang ekstensif, mereka adalah statistik sampel yang merupakan fungsi dari data yang mendasarinya. Dengan demikian, mereka tunduk pada variabilitas pengambilan sampel untuk alasan yang sama dan dengan cara yang sama seperti statistik biasa, dan karenanya memiliki distribusi pengambilan sampel yang mencirikan keakuratan perkiraan. Untuk ukuran sampel yang cukup besar, distribusi sampel ini mendekati Gaussian, dan distribusi sampling gabungan dari parameter yang diestimasi secara simultan mendekati multivariat Gaussian (misalnya, distribusi sampling dari estimasi α dan β dalam Persamaan 2.73 akan mendekati normal bivariat).

Diberikan $\theta = [\theta_1, \theta_2, \dots, \theta_k]$ menjadi vektor parameter berdimensi k yang akan diestimasi. Misalnya pada persamaan 4,73 $k = 2$, $\theta_1 = a$ dan $\theta_2 = \beta$. Estimasi matriks varians-kovarians untuk distribusi sampling Gaussian multivariat ([∑], dalam Persamaan 10.1) diberikan oleh invers dari matriks informasi yang diboboti dengan estimasi nilai parameter $\hat{\theta}$.

$$\widehat{Var}(\hat{\theta}) = \left[I(\hat{\theta})\right]^{-1} \tag{3.77}$$

(Notasi aljabar matriks dijelaskan di Bab 9). Matriks informasi pada gilirannya dihitung dari turunan kedua dari fungsi log-likelihood sehubungan dengan vektor parameter dan dievaluasi dengan nilai perkiraannya.

$$\left[I(\hat{\theta})\right] = -\begin{bmatrix} \dfrac{\partial^2 L}{\partial \hat{\theta}_1^2} & \dfrac{\partial^2 L}{\partial \hat{\theta}_1 \partial \hat{\theta}_2} & \cdots & \dfrac{\partial^2 L}{\partial \hat{\theta}_1 \partial \hat{\theta}_k} \\ \dfrac{\partial^2 L}{\partial \hat{\theta}_2 \partial \hat{\theta}_1} & \dfrac{\partial^2 L}{\partial \hat{\theta}_2^2} & \cdots & \dfrac{\partial^2 L}{\partial \hat{\theta}_2 \partial \hat{\theta}_k} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 L}{\partial \hat{\theta}_k \partial \hat{\theta}_1} & \dfrac{\partial^2 L}{\partial \hat{\theta}_k \partial \hat{\theta}_2} & \cdots & \dfrac{\partial^2 L}{\partial \hat{\theta}_k^2} \end{bmatrix} \tag{3.78}$$

Perhatikan bahwa invers matriks informasi muncul sebagai bagian dari iterasi Newton-Raphson untuk estimasi itu sendiri, misalnya untuk estimasi parameter distribusi gamma pada Persamaan 2.73. Keuntungan menggunakan algoritme ini adalah estimasi varians dan kovarians untuk distribusi sampling gabungan untuk parameter estimasi telah dihitung

pada iterasi terakhir. Algoritme EM tidak memberikan besaran ini secara otomatis, tetapi tentu saja dapat dihitung dari parameter yang diestimasi; baik dengan mengganti estimasi parameter ke dalam ekspresi analitik untuk turunan kedua dari fungsi log-likelihood atau dengan perkiraan perbedaan hingga dari turunannya.

3.7 Simulasi Statistik

Tema mendasar dari bab ini adalah bahwa ketidakpastian dalam proses fisik dapat dijelaskan dengan distribusi probabilitas yang sesuai. Ketika komponen dari fenomena fisik atau proses yang menarik tidak pasti, fenomena atau proses tersebut masih dapat dipelajari melalui simulasi komputer, menggunakan algoritma yang menghasilkan angka yang dapat dilihat sebagai sampel acak dari distribusi probabilitas yang relevan. Generasi dari angka-angka yang tampaknya acak ini disebut simulasi statistik.

Bagian ini menjelaskan algoritma yang digunakan dalam simulasi statistik. Algoritme ini terdiri dari fungsi rekursif deterministik, sehingga keluarannya tidak benar-benar acak sama sekali. Faktanya, keluarannya dapat diduplikasi persis jika diinginkan, yang dapat berguna saat men-debug kode dan menjalankan replikasi terkontrol dari eksperimen numerik. Meskipun algoritme ini kadang-kadang disebut generator bilangan acak, nama yang lebih tepat adalah pembangkit bilangan acak semu karena keluaran deterministiknya hanya

tampak acak. Namun, hasil yang cukup berguna dapat diperoleh dengan menganggapnya acak secara efektif.

Pada dasarnya, setiap pembangkitan bilangan acak dimulai dengan simulasi dari distribusi seragam, dengan PDF $f(u) = 1,\ 0 \leq u \leq 1$, yang dijelaskan pada Bagian 4.7.1. Mensimulasikan nilai dari distribusi lain melibatkan transformasi satu atau lebih variabel seragam. Lebih banyak tentang subjek daripada yang dapat disajikan di sini, termasuk kode dan kode semu untuk banyak algoritma tertentu, dapat ditemukan dalam referensi seperti Boswell et al. (1993), Bratley et al. (1987), Dagpunar (1988), Press et al. (1986), Tezuka (1995), dan ensiklopedia Devroye (1986).

Materi pada bagian ini berkaitan dengan pembangkitan variabel acak bebas skalar. Pembahasan menekankan pembuatan variabel kontinu, tetapi dua metode umum yang dijelaskan pada Bagian 4.7.2 dan 4.7.3 juga dapat digunakan untuk distribusi diskrit. Perluasan simulasi statistik untuk urutan yang berkorelasi disertakan dalam bagian 8.2.4 dan 8.3.7 pada model deret waktu domain waktu. Ekstensi untuk simulasi multivariat disajikan pada Bagian 10.4.

3.7.1 Generator Angka Acak Seragam

Seperti disebutkan sebelumnya, simulasi statistik bergantung pada ketersediaan algoritme yang baik untuk menghasilkan sampel acak dan tidak berkorelasi dari distribusi seragam (0, 1) yang dapat diubah untuk mensimulasikan sampel acak dari

distribusi lain. Secara aritmatika, generator bilangan acak seragam mengambil nilai bilangan bulat awal yang disebut seed, memprosesnya untuk menghasilkan nilai seed yang diperbarui, lalu menskalakan seed yang diperbarui ke interval (0, 1). Benih awal dipilih oleh pemrogram, tetapi biasanya pemanggilan berikutnya ke algoritma pembangkitan seragam beroperasi pada benih yang paling baru diperbarui. Operasi aritmatika yang dilakukan oleh algoritme sepenuhnya deterministik, jadi memulai ulang generator dengan benih yang disimpan sebelumnya memungkinkan reproduksi yang tepat dari urutan angka "acak" yang dihasilkan.

Algoritma yang paling umum ditemui untuk menghasilkan bilangan acak seragam adalah generator kongruensi linier, yang didefinisikan oleh:

$$S_n = a\, s_{n-1} + c, \text{mod M} \qquad (3.79a)$$

dan

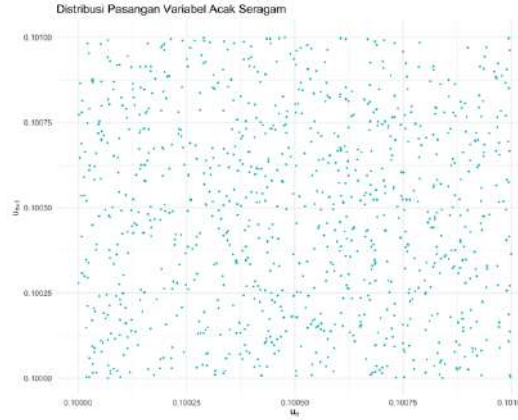
$$u_n = \frac{s_n}{M} \qquad (3.79b)$$

Di sini S_{n-1} adalah nilai benih yang dimajukan dari iterasi sebelumnya, S_n adalah nilai benih yang diperbarui, dan a, c, dan M adalah parameter bilangan bulat yang masing-masing disebut pengali, kenaikan, dan modulus. U_n dalam Persamaan 2.79b adalah variabel acak seragam yang dihasilkan oleh iterasi yang didefinisikan oleh Persamaan 2.79. Karena benih yang diperbarui S_n adalah sisa ketika $aS_{n-1} + c$ dibagi

dengan M, S_n harus kurang dari M, dan hasil bagi dalam Persamaan 2.79b akan kurang dari 1. Untuk a > 0 dan $c \geq 0$, Persamaan 2.79b harus lebih besar dari 0. Parameter dalam Persamaan 2.79a harus dipilih dengan hati-hati jika generator kongruensi linier ingin bekerja dengan baik. Urutan S_n berulang dengan periode paling banyak M, dan umum untuk memilih modulus sebagai bilangan prima hampir sebesar bilangan bulat terbesar yang dapat diwakili oleh komputer yang menjalankan algoritme. Banyak komputer menggunakan bilangan bulat 32-bit (yaitu 4-byte), dan $M = 2^{31} - 1$ adalah pilihan umum, seringkali dikombinasikan dengan $a = 16807$ dan $c = 0$.

Generator kongruen linier dapat cocok untuk beberapa tujuan, terutama dalam aplikasi dimensi rendah. Namun, dalam dimensi yang lebih tinggi, hasilnya tidak berpola mengisi ruang. Secara khusus, pasangan u berurutan dari Persamaan 2.79b jatuh pada rangkaian garis sejajar pada bidang $u_n - u_{n+1}$, tiga kali lipat u berturut-turut dari Persamaan 2.79b jatuh pada rangkaian bidang sejajar dalam volume yang ditentukan oleh sumbu $u_n - u_{n+1} - u_{n+2}$, dan seterusnya dengan jumlah kenampakan paralel ini berkurang dengan cepat dengan meningkatnya dimensi k, kira-kira menurut $(k!\,M)^{\frac{1}{k}}$. Inilah alasan lain untuk Memilih modul M sebesar mungkin, karena untuk $M = 2^{31} - 1$ dan $k = 2$, $(k!\,M)^{\frac{1}{k}}$ adalah sekitar 65.000.

Gambar 3.17 menunjukkan tampilan yang diperbesar dari sebagian bujur sangkar satuan dimana 1000 pasang bilangan acak seragam yang tidak tumpang tindih yang dihasilkan menggunakan Persamaan 2.79 telah diplot. Domain kecil ini berisi 17 garis paralel yang jatuh secara berurutan dari generator ini, dengan jarak interval 0,000059. Perhatikan bahwa jarak minimum titik-titik vertikal jauh lebih sempit, yang menunjukkan bahwa jarak garis titik-titik hampir vertikal tidak menentukan resolusi generator. Jarak horizontal yang relatif sempit pada **Gambar 3.17** menunjukkan bahwa generator kongruensi linier sederhana mungkin tidak terlalu kasar untuk beberapa tujuan dimensi rendah (walaupun lihat Bagian 4.7.4 untuk interaksi patologis dengan algoritma umum untuk menghasilkan variabel Gaussian dalam dua dimensi). Namun, dalam dimensi yang lebih tinggi, jumlah hyperplane yang kelompok nilai berturut-turut dari generator kongruen linier dibatasi



Gambar 3.17 1000 pasangan variabel acak seragam yang tidak tumpang tindih di bagian kecil persegi yang didefinisikan oleh $0 < u_n < 1$ dan $0 < u_{n+1} < 1$; dihasilkan menggunakan Persamaan 2.79, dan $a = 16807$, $c = 0$ dan $M = 2^{31} - 1$. Wilayah kecil ini berisi 17 dari sekitar 65.000 garis sejajar dimana pasangan berurutan jatuh di seluruh unit persegi.

menurun dengan cepat, membuat algoritme semacam ini tidak mungkin menghasilkan banyak kombinasi yang seharusnya mungkin: Untuk k = 3, 5, 10, dan 20-dimensi adalah jumlah hyperplane, yang semuanya dikatakan poin yang dihasilkan secara acak termasuk, masing-masing lebih kecil dari 2350, 200, 40 dan 25, bahkan untuk modul yang relatif besar $M = 2^{31} - 1$. Perhatikan bahwa situasinya bisa jauh lebih buruk jika parameternya dipilih dengan buruk: generator terkenal tetapi sebelumnya tersebar luas yang disebut RANDU (Persamaan 2.79 dengan a = 65539, $c = 0$ dan $M = 2^{31}$) dibatasi hanya 15 level dalam tiga dimensi.

Penggunaan langsung generator seragam kongruen linier tidak dapat direkomendasikan karena hasil berpola mereka dalam dua dimensi atau lebih. Algoritme yang lebih baik dapat dibangun dengan menggabungkan dua atau lebih generator kongruen linier yang berjalan secara independen, atau dengan menggunakan satu generator tersebut untuk menggabungkan keluaran dari yang lain; Contohnya ada di Bratley et al. (1987) dan Press et al. (1986). Alternatif yang menarik dengan sifat yang tampaknya sangat bagus adalah algoritma yang relatif baru yang disebut Mersenne Twister (Matsumoto dan Nishimura 1998), yang tersedia secara gratis

dan dapat dengan mudah ditemukan dengan mencari di web untuk nama itu.

3.7.2 Pembuatan bilangan acak tidak seragam dengan inversi

Pembalikan adalah metode pembuatan variabel tidak seragam yang paling mudah untuk dipahami dan diprogram ketika fungsi kuantil $F^{-1}(p)$ (Persamaan 2.19) dalam bentuk tertutup. Oleh karena itu terlepas dari bentuk fungsional CDF $F(x)$, distribusi variabel $u = F(x)$ yang didefinisikan oleh transformasi ini pada [0, 1] adalah seragam. Kebalikannya juga benar, sehingga CDF dari variabel yang ditransformasikan $x(F) = F^{-1}(u)$ sama dengan $F(x)$, dimana distribusi u pada [0, 1] adalah seragam. Oleh karena itu, untuk membangkitkan variabel acak dengan CDF F(x) yang fungsi kuantilnya dalam bentuk tertutup, kita hanya perlu membangkitkan variabel acak seragam seperti dijelaskan pada Bagian 4.7.1 dan membalikkan CDF dengan mengganti nilai ini dengan nilai yang sesuai. fungsi kuantil nilai.

Inversi juga dapat digunakan untuk distribusi bentuk tertutup tanpa fungsi kuantil dengan menggunakan perkiraan numerik, evaluasi iteratif, atau pencarian tabel interpolasi. Namun, bergantung pada distribusinya, solusi ini mungkin tidak cukup cepat atau akurat, sehingga metode lain akan lebih sesuai.

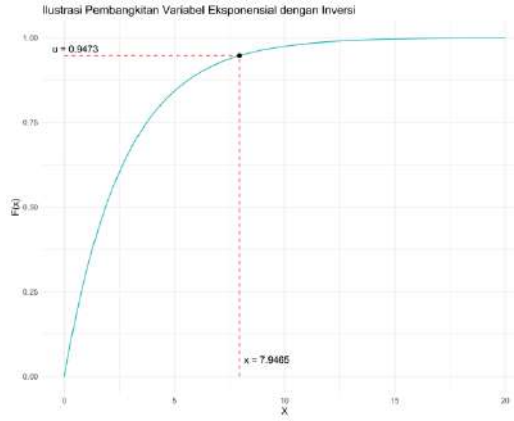
Contoh 3.13 Pembuatan variabel eksponensial menggunakan inversi

Distribusi eksponensial (Persamaan 2.45 dan 4.46) adalah distribusi kontinu sederhana yang fungsi kuantilnya ada dalam bentuk tertutup. Secara khusus, memecahkan Persamaan 2.46 untuk probabilitas kumulatif p menghasilkan

$$F^{-1}(p) = -\beta \ln(1-p) \tag{3.80}$$

Menghasilkan variabel acak yang terdistribusi secara eksponensial hanya perlu mengganti probabilitas kumulatif p dalam Persamaan 2.80 dengan variabel acak yang seragam, yaitu $x(F) = F^{-1}(u) = -\beta \ln(1-u)$. **Gambar 3.18** mengilustrasikan proses untuk u yang dipilih secara sewenang-wenang dan distribusi eksponensial dengan rata-rata $\beta = 2.7$. Perhatikan bahwa nilai numerik pada **Gambar 3.18** telah dibulatkan menjadi beberapa digit signifikan untuk kenyamanan, tetapi dalam praktiknya semuanya signifikan Digit akan dipertahankan dalam perhitungan.

Karena distribusi seragam simetris dengan rata-rata 0,5, distribusi $1-u$ adalah distribusi seragam yang sama dengan distribusi u, sehingga variasi eksponensial dimungkinkan.



Gambar 3.18 Ilustrasi pembangkitan variabel eksponensial dengan inversi. Kurva halus adalah CDF (Persamaan 2.46) dengan rata-rata $\beta = 2.7$. Variabel seragam $u = 0.9473$ diubah menjadi variabel eksponensial yang dihasilkan $x = 7.9465$ dengan kebalikan dari CDF. Gambar ini juga mengilustrasikan bahwa inversi menghasilkan transformasi monoton dari varian seragam yang mendasarinya.

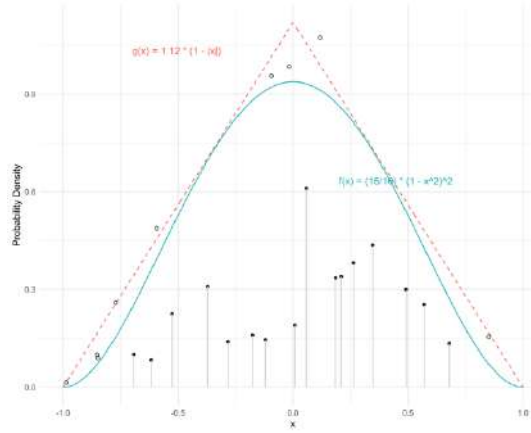
Dapat hitung dengan mudah $x(F) = F^{-1}(1-u) = -\beta \ln(u)$. Bahkan jika ini sedikit lebih mudah secara komputasi, masih layak menggunakan $-\beta \ln(1-u)$ untuk menjaga kemonotonan metode inversi; yaitu bahwa kuantil dari distribusi seragam yang mendasarinya sesuai persis dengan kuantil distribusi variabel acak yang dihasilkan, sehingga u terkecil sesuai dengan x terkecil dan u terbesar dengan x terbesar. Satu kasus dimana properti ini dapat berguna adalah saat membandingkan simulasi yang mungkin bergantung pada parameter berbeda atau distribusi berbeda. Mempertahankan monoton atas kumpulan simulasi semacam itu (dan memulai masing-masing dengan seed nomor acak

yang sama) dapat memungkinkan perbandingan yang lebih akurat antara simulasi yang berbeda, karena proporsi varians yang lebih besar dalam perbedaan antara simulasi kemudian disebabkan oleh perbedaan dalam simulasi. proses. dan lebih sedikit karena variasi pengambilan sampel dalam aliran nomor acak. Teknik ini dikenal dalam literatur simulasi sebagai pengurangan varians.

3.7.3 Pembangkitan Bilangan Acak Tidak Seragam dengan Penolakan

Metode inversi nyaman secara matematis dan komputasi ketika fungsi kuantil dapat dengan mudah dievaluasi, tetapi sebaliknya dapat menjadi rumit. Pendekatan yang lebih umum adalah metode penolakan, atau metode penerimaan-penolakan, yang hanya mensyaratkan bahwa PDF $f(x)$ dari distribusi yang akan disimulasikan dapat dievaluasi secara eksplisit. Selain itu, bagaimanapun juga, sebuah amplop PDF $g(x)$ harus ditemukan. Kerapatan selubung $g(x)$ harus memiliki dukungan yang sama dengan $f(x)$ dan harus mudah disimulasikan (misalnya dengan inversi). Sebagai tambahan, sebuah konstanta c harus ditemukan sehingga $f(x) \leq cg(x)$, untuk semua x dengan probabilitas bukan nol. Artinya, $f(x)$ harus didominasi oleh fungsi $cg(x)$ untuk semua x yang relevan. Bagian tersulit dari merancang algoritma penolakan adalah menemukan amplop PDF yang cocok dengan bentuk yang mirip dengan distribusi yang akan disimulasikan, sehingga konstanta c dapat sedekat mungkin dengan 1.

Setelah amplop PDF dan konstanta c yang cukup untuk memastikan dominasi telah ditemukan, simulasi berlanjut melalui penolakan dalam dua langkah, yang masing-masing memerlukan pemanggilan independen ke generator seragam. Pertama, varian kandidat dihasilkan dari $g(x)$ menggunakan:



Gambar 3.19 Representasi simulasi dari kerapatan kuartik (biweight), $f(x) = \left(\frac{15}{16}\right)(1 - x2)^2$ (Tabel 1.1), menggunakan kerapatan segitiga (Tabel 1.1) sebagai amplop, dengan $c = 1.12$. Dua puluh lima kandidat x disimulasikan dari kerapatan segitiga, 21 di antaranya diterima ( ) karena mereka juga berada di bawah distribusi $f(x)$ yang akan disimulasikan, dan 4 ditolak (O) karena berada di luarnya. Garis abu-abu terang menunjukkan nilai yang disimulasikan pada sumbu horizontal.

Variabel unik pertama u_1, mungkin dengan inversi sebagai $x = G^{-1}(u_1)$. Kedua, kandidat x dikenai uji acak menggunakan varian seragam kedua: kandidat x diterima jika

$u_2 \leq \frac{f(x)}{[cg(x)]}$, jika tidak kandidat x ditolak dan metode dicoba lagi dengan sepasang varian unik yang baru.

Gambar 3.19 mengilustrasikan metode penolakan untuk mensimulasikannya dari kerapatan quart (lihat Tabel 1.1). PDF untuk distribusi ini adalah polinomial tingkat empat, sehingga CDF-nya dapat dengan mudah ditemukan sebagai polinomial tingkat lima melalui integrasi. Namun, secara eksplisit membalikkan CDF (menyelesaikan polinomial derajat kelima) bisa menjadi masalah, jadi penolakan adalah cara yang masuk akal untuk mensimulasikan distribusi ini. Distribusi segitiga (juga diberikan pada Tabel 1.1) dipilih sebagai distribusi amplop $g(x)$; dan konstanta $c = 1.12$ cukup untuk $cg(x)$ mendominasi $f(x)$ pasa $-1 \leq x \leq 1$. Fungsi segitiga merupakan pilihan yang masuk akal untuk densitas lambung kapal karena mendominasi $f(x)$ dengan nilai konstanta regang c yang relatif kecil, sehingga kemungkinan calon x ditolak relatif kecil. Selain itu, cukup sederhana sehingga kita dapat dengan mudah menurunkan fungsi kuantilnya, memungkinkan simulasi dengan inversi. Secara khusus, mengintegrasikan PDF segitiga memberikan CDF.

$$G(x) = \begin{cases} \frac{x^2}{2} + x + \frac{1}{2}, & -1 \leq x \leq 0, \quad (3.81a) \\ -\frac{x^2}{2} + x + \frac{1}{2}, & 0 \leq x \leq 1, \quad (3.81b) \end{cases}$$

yang dapat dibalik untuk mendapatkan fungsi kuantil

$$x(G) = G^{-1}(p) = \begin{cases} \sqrt{2p} - 1, & 0 \le p \le \frac{1}{2} \quad (3.82a) \\ 1 - \sqrt{2(1-p)}, & \frac{1}{2} \le p \le 1 \quad (3.82b) \end{cases}$$

Gambar 3.19 menunjukkan 25 titik kandidat, 21 di antaranya diterima (X), dengan garis abu-abu terang menunjuk ke nilai yang dihasilkan sesuai pada sumbu horizontal. Itu koordinat horizontal dari titik-titik ini adalah $G^{-1}(u_1)$; yaitu, penarikan acak dari kernel segitiga $g(x)$ menggunakan variabel seragam u_1. Koordinat vertikalnya adalah $u_2 cg[G^{-1}(u_1)]$, yang merupakan jarak merata antara sumbu horizontal dan $cg(x)$. Ditentukan dengan menggunakan variabel acak seragam kedua u_2 pada kandidat x. Intinya, algoritma penolakan berfungsi karena dua variabel seragam menentukan titik yang terdistribusi secara merata (dalam dua dimensi) di bawah fungsi $cg(x)$, dan kandidat x diterima sesuai dengan probabilitas bersyarat yang juga berada di bawah PDF $f(x)$ kebohongan. Metode penolakan dengan demikian sangat mirip dengan integrasi Monte Carlo dari $f(x)$. Ilustrasi simulasi distribusi ini dengan penolakan diberikan pada Contoh 4.15.

Kerugian dari metode penolakan adalah bahwa beberapa pasangan variabel seragam terbuang sia-sia ketika kandidat x ditolak, dan untuk alasan ini diharapkan konstanta c sekecil mungkin: probabilitas kandidat x akan ditolak adalah $1 - 1/c (= 0.107$ untuk situasi di **Gambar 3.19**). Properti lain dari metode ini adalah bahwa nomor acak tak tentu dari

varian seragam diperlukan untuk satu pemanggilan algoritme, sehingga sinkronisasi aliran nomor acak yang dimungkinkan oleh metode inversi paling sulit dicapai saat menggunakan penolakan.

3.7.4 Metode Box-Muller untuk membangkitkan bilangan acak Gaussian

Salah satu distribusi yang paling sering dibutuhkan dalam simulasi adalah distribusi Gaussian (Persamaan 2.23). Karena CDF untuk distribusi ini tidak ada dalam bentuk tertutup, fungsi kuantilnya juga tidak ada, sehingga pembuatan variabel Gaussian dengan inversi hanya dapat dilakukan secara kira-kira. Alternatifnya, variabel Gaussian standar (Persamaan 2.24) dapat dihasilkan berpasangan menggunakan transformasi pintar dari sepasang variabel seragam independen dengan algoritma yang disebut metode Box-Muller. Dimensi yang sesuai (non-standar) variabel Gaussian kemudian dapat direkonstruksi menggunakan rata-rata distribusi dan varians menurut Persamaan 2.28.

Metode Box-Muller menghasilkan pasangan variabel normal bivariat standar independen z_1 dan z_2 yaitu sampel acak dari PDF bivariat pada Persamaan 2.36 - dengan korelasi $\rho = 0$, sehingga kontur bidang PDF adalah lingkaran. Karena kontur bidang adalah lingkaran, setiap arah yang jauh dari titik asal memiliki kemungkinan yang sama, sehingga dalam koordinat kutub, PDF untuk sudut titik acak adalah seragam pada [0, 21r]. Sudut seragam dalam interval ini dapat dengan mudah

disimulasikan sebagai $\theta = 2\pi u_1$ dari yang pertama dari pasangan varian seragam independen. CDF untuk jarak radial dari Gaussian bivariat standar adalah

$$F(r) = 1 - \exp\left[-\frac{r^2}{2}\right], 0 \leq r \leq \infty \tag{3.83}$$

yang dikenal sebagai distribusi Rayleigh. Persamaan 2.83 dapat dengan mudah dibalik untuk mendapatkan fungsi kuantil $r(F) = F^{-1}(u_2) = -2\ln(1 - u_2)$. Ketika diubah kembali ke koordinat Cartesian, pasangan variabel Gaussian standar independen yang dihasilkan adalah

$$z_1 = \cos(2\pi\mu_1)\sqrt{-2\ln(\mu_2)} \tag{3.84a}$$

$$z_2 = \sin(2\pi\mu_1)\sqrt{-2\ln(\mu_2)} \tag{3.84b}$$

Metode Box-Muller sangat umum dan populer, tetapi harus berhati-hati saat memilih generator yang seragam untuk mengoperasikannya. Secara khusus, garis-garis pada bidang $u_1 - u_2$ dihasilkan oleh generator kongruensi linier sederhana, seperti yang ditunjukkan pada **Gambar 3.17**, diproses oleh transformasi polar pada Persamaan 2.84 untuk menghasilkan spiral pada bidang $z_1 - z_2$, seperti yang dijelaskan lebih lengkap oleh Bratley et al. (1987). Pemolaan ini jelas tidak diinginkan, dan generator seragam yang lebih canggih sangat penting saat membuat variabel Box-Muller-Gaussian.

3.7.5 Simulasi Distribusi Campuran Dan Perkiraan Densitas Kernel

Simulasi dari distribusi campuran (Persamaan 2.63) hanya sedikit lebih rumit daripada simulasi dari salah satu komponen PDF. Ini adalah prosedur dua langkah dimana distribusi komponen dipilih berdasarkan bobot w, yang dapat dilihat sebagai probabilitas distribusi komponen yang dipilih. Setelah distribusi komponen dipilih secara acak, varian dari distribusi tersebut dihasilkan dan dikembalikan sebagai sampel simulasi dari campuran.

Misalnya, pertimbangkan simulasi dari distribusi eksponensial campuran, Persamaan 2.66, yang merupakan campuran probabilitas dari dua PDF eksponensial. Dua variabel acak seragam independen diperlukan untuk menghasilkan realisasi dari distribusi ini: satu variabel acak seragam untuk memilih salah satu dari dua distribusi eksponensial dan yang lainnya untuk disimulasikan dari distribusi ini. Menggunakan inversi untuk langkah kedua (Persamaan 2.80), prosedurnya sederhana

$$x = \begin{cases} -\beta_1 \ln(1-\mu_2), & \mu_1 \le w \quad (3.85a) \\ -\beta_2 \ln(1-\mu_2), & \mu_1 > w \quad (3.85b) \end{cases}$$

Di sini distribusi eksponensial dengan rata-rata $\beta = 1$ dengan probabilitas w dipilih menggunakan u_1; dan kebalikan dari salah satu dari dua distribusi yang dipilih diimplementasikan menggunakan variabel seragam kedua u_2.

Estimasi densitas kernel yang dijelaskan pada Bagian 3.3.6 adalah contoh yang menarik dari distribusi campuran. Di sini campuran terdiri dari n PDF yang kemungkinannya sama, masing-masing sesuai dengan salah satu dari n pengamatan variabel x. PDF ini sering menggunakan salah satu formulir yang tercantum dalam Tabel 1.1. Sekali lagi, langkah pertama adalah memilih n nilai data mana yang menjadi pusat kernel yang akan disimulasikan pada langkah kedua, yang dapat dilakukan seperti ini:

$$\text{memilih } x_i \text{ jika } \frac{i-1}{n} \leq u < \frac{i}{n} \tag{3.86a}$$

dengan

$$i = int[n\,u + 1] \tag{3.86b}$$

Di sini int[•] hanya menunjukkan mempertahankan bagian bilangan bulat atau memotong pecahan.

Contoh 3.14 Simulasi dari estimasi densitas kernel pada Gambar 1.8b

Gambar 1.8b menunjukkan perkiraan densitas kernel yang mewakili data suhu Guayaquil pada Tabel A.3; dibangun menggunakan Persamaan 1.13, kernel kuartik (lihat Tabel 1.1) dan parameter pemulusan $h = 0.6$. Menggunakan penolakan untuk mensimulasikan dari kerapatan kernel kuartik, setidaknya tiga varian seragam independen diperlukan untuk mensimulasikan sampel acak dari distribusi

ini. Misalkan ketiga variabel seragam ini dihasilkan sebagai $u_1 = 0.257990, u_2 = 0.898875$, dan $u_3 = 0\ 465617$.

Langkah pertama adalah memilih mana dari $n = 20$ nilai suhu pada Tabel A.3 yang akan digunakan untuk memusatkan kernel yang akan disimulasikan. Menggunakan Persamaan 2.86b, ini adalah x_i dimana $i = int[20 \bullet 0.257990 + 1] = int[6.1598] = 6$ menghasilkan $T_i = 24.3$℃ karena $i = 6$ sesuai dengan tahun 1956.

Langkah kedua adalah simulasi dari kuartik kernel, yang dapat dilakukan dengan penolakan, seperti yang ditunjukkan pada **Gambar 3.19**. Pertama, kandidat x dihasilkan dari distribusi segitiga dominan dengan inversi (Persamaan 2.82b) menggunakan variabel seragam kedua, $u_2 = 0.898875$. Perhitungan ini menghasilkan $x(G) = 1 - [2(1 - 0.898875)]^{\frac{1}{2}} = 0.550278$. Apakah nilai ini diterima atau ditolak? Pertanyaan ini dijawab dengan membandingkan u_3 dengan rasio $\frac{f(x)}{[cg(x)]}$, dimana $f(x)$ adalah PDF kuartik, $g(x)$ adalah PDF segitiga, dan $c = 1.12$ agar $cg(x)$ untuk mendominasi $f(x)$. Kami kemudian menemukan bahwa $u_3 = 0.465617 < \frac{0.455700}{[1.12 \bullet 0.449722]} = 0.904726$, sehingga kandidat $x = 0.550278$ diterima.

Nilai x yang baru saja dihasilkan adalah undian acak dari kernel quartic nol-pusat standar dengan parameter pemulusan kesatuan. Menyamakannya dengan argumen fungsi kernel K pada Persamaan 1.13 memberikan

$x = 0.550278 = \frac{T-T_i}{h} = \frac{T-24.3℃}{0.6}$ yang memusatkan kernel pada T_i dan menskalakan sesuai sehingga hasil akhir mensimulasikan nilai T = (0.550278)(0.6) + 24.3 = 24.6℃.

Latihan

1. Dengan menggunakan distribusi binomial sebagai model untuk membekukan Danau Cayuga, seperti yang ditunjukkan pada Contoh 1.1 dan 1.2, hitung probabilitas bahwa danau tersebut akan membeku setidaknya sekali selama empat tahun.

2. Hitung probabilitas bahwa Danau Cayuga akan membeku selanjutnya
   a. Tepat 5 tahun.
   b. Dalam 25 tahun atau lebih.

3. Dalam sebuah artikel yang diterbitkan dalam jurnal Science, Gray (1990) membandingkan berbagai aspek badai Atlantik yang terjadi selama musim kemarau dan musim hujan di Afrika sub-Sahara. Selama 18 tahun kekeringan tahun 1970–1987, hanya satu badai besar (Intensitas 3 atau lebih besar) yang terjadi di pantai timur Amerika Serikat, tetapi 13 badai seperti itu melanda Amerika Serikat bagian timur selama musim hujan 23 tahun pada tahun 1947. –1969.
   a. Asumsikan bahwa jumlah angin topan yang terjadi di AS bagian timur mengikuti distribusi Poisson yang sifat-sifatnya bergantung pada curah hujan Afrika. Tetapkan dua distribusi Poisson dengan data Gray (satu karena kekeringan dan satu lagi karena musim hujan di Afrika Barat).
   b. Mengingat tahun kering di Afrika Barat, hitung

probabilitas bahwa setidaknya satu badai besar akan melanda Amerika Serikat bagian timur.

c. Mengingat tahun basah di Afrika Barat, hitung probabilitas bahwa setidaknya satu badai besar akan melanda Amerika Serikat bagian timur.

4. Misalkan rata-rata badai yang kuat membuat pendaratan di Amerika Serikat bagian timur kerusakan $5 miliar. Berapa nilai yang diharapkan dari kerusakan badai tahunan dari badai tersebut menurut masing-masing dari dua distribusi bersyarat dalam nomor 3?

5. Menggunakan data suhu bulan Juni untuk Guayaquil, Ekuador pada Tabel A.3,
   a. Pasang distribusi Gaussian.
   b. Tanpa mengubah setiap nilai data, tentukan dua parameter Gaussian yang akan dihasilkan jika data ini dinyatakan dalam °F.
   c. Buat histogram dari data suhu ini dan lapisi fungsi densitasnya distribusi pas dalam histogram.

6. Menggunakan distribusi Gaussian dengan $\mu = 19$°C dan $\alpha = 1\ 7$°C:
   a. Perkirakan probabilitas bahwa suhu di bulan Januari (untuk Miami, Florida) akan lebih dingin dari 15°C.
   b. Berapa suhu yang akan lebih tinggi dari semuanya kecuali 1% terhangat di bulan Januari di Miami?

7. Untuk data curah hujan bulan Juli untuk Ithaca disajikan pada Tabel 1.9
    a. Sesuaikan distribusi gamma dengan estimator kemungkinan maksimum menggunakan pendekatan Thom.
    b. Tanpa mengonversi setiap nilai data, temukan nilai dari kedua parameter yang dihasilkan jika data dinyatakan dalam mm.
    c. Buatlah histogram dari data presipitasi ini dan lapisi dengan fungsi densitas gamma yang sesuai.

8. Gunakan hasil dari nomor 7 untuk menghitung:
    a. Persentil ke-30 dan ke-70 dari presipitasi bulan Juli di Ithaca.
    b. Perbedaan antara rata-rata sampel dan median distribusi.
    c. Probabilitas curah hujan Ithaca setidaknya 7 inci di bulan Juli mendatang.

9. Hitung ulang nomor 8 menggunakan distribusi log-normal untuk memplot data pada Tabel 1.9.

10. Rata-rata kedalaman salju maksimum untuk setiap musim dingin di lokasi yang diinginkan adalah 80 cm, dan standar deviasi (mencerminkan perbedaan tahunan dalam kedalaman salju maksimum) adalah 45 cm.
    a. Sesuaikan distribusi Gumbel untuk memplot data ini menggunakan metode momen.
    b. Turunkan fungsi kuantil untuk distribusi Gumbel dan

gunakan untuk memperkirakan kedalaman salju yang dilampaui rata-rata hanya dalam satu tahun dalam 100.

DAFTAR PUSTAKA

Abramowitz, M., and I.A. Stegun, eds., 1984. *Pocketbook of Mathematical Functions. Frankfurt*, Verlag Harri Deutsch, 468 pp.

Agresti, A., 1996. *An Introduction to Categorical Data Analysis*. Wiley, 290 pp. Agresti, A., and B.A. Coull, 1998. Approximate is better than "exact" for interval estimation of binomial proportions. American Statistician, 52, 119–126.

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

Allen, M.R., and L.A. Smith, 1996. Monte Carlo SSA: Detecting irregular oscillations in the presence of colored noise. *Journal of Climate*, 9, 3373–3404.

Anderson, J.L., 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9, 1518–1530.

Anderson, J.L., 1997. The impact of dynamical constraints on the selection of initial conditions for ensemble predictions: low-order perfect model results. *Monthly Weather Review*, 125, 2969–2983.

Buizza, R., M. Miller, and T.N. Palmer, 1999b. Stochastic representation of model uncer- tainties in the ECMWF Ensemble Prediction System. *Quarterly Journal of the Royal Meteorological Society*, 125, 2887–2908.

Burman, P., E. Chow, and D. Nolan, 1994. A cross-validatory method for dependent data. *Biometrika*, 81, 351–358.

Carter, G.M., J.P. Dallavalle, and H.R. Glahn, 1989. Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Weather and Forecasting*, 4, 401–412.

Cunnane, C., 1978. Unbiased plotting positionsa review. *Journal of Hydrology*, 37, 205–222.

Daan, H., 1985. Sensitivity of verification scores to the classification of the predictand. *Monthly Weather Review*, 113, 1384–1392.

Ehrendorfer, M., 1997. *Predicting the uncertainty of numerical weather forecasts: a review*. Meteorol. Zeitschrift, 6, 147–183.

Gleeson, T.A., 1970. Statistical-dynamical predictions. *Journal of Applied Meteorology*, 9, 333–344.

Goldsmith, B.S., 1990. *NWS verification of precipitation type and snow amount forecasts during the AFOS era*. NOAA Technical Memorandum NWS FCST 33. National Weather Service, 28 pp.

Golub, G.H., and C.F. van Loan, 1996. *Matrix Computations. Johns Hopkins Press*, 694 pp.

Houtekamer, P.L., L. Lefaivre, J. Derome, H. Ritchie, and H.L. Mitchell, 1996. A system simulation approach to ensemble prediction. *Monthly Weather Review*, 124, 1225–1242.

Houtekamer, P.L., and H.L. Mitchell, 1998. Data assimilation using an ensemble Kalman filtering technique. *Monthly Weather Review*, 126, 796–811.

Hsu, W.-R., and A.H. Murphy, 1986. The attributes diagram: a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, 2, 285–293.

Hu, Q., 1997. On the uniqueness of the singular value decomposition in meteorological applications. *Journal of Climate*, 10, 1762–1766.

Klein, W.H., B.M. Lewis, and I. Enger, 1959. Objective prediction of five-day mean temperature during winter. *Journal of Meteorology*, 16, 672–682.

Knaff, J.A., and C.W. Landsea, 1997. An El Niño-southern oscillation climatology and persistence (CLIPER) forecasting scheme. *Weather and Forecasting*, 12, 633–647.

Krzysztofowicz, R., 1983. *Why should a forecaster and a decision maker use Bayes' theorem? Water Resources Research*, 19, 327–336.

Krzysztofowicz, R., 2001. The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, 249, 2–9.

Lin, J. W.-B., and J.D. Neelin, 2002. *Considerations for stochastic convective parameterization. Journal of the Atmospheric Sciences, 59, 959–975. Lindgren, B.W., 1976*. Statistical Theory. MacMillan, 614 pp.

Lipschutz, S., 1968. *Schaum's Outline of Theory and Problems of Linear Algebra*. McGraw-Hill, 334 pp.

Livezey, R.E., 1985. Statistical analysis of general circulation model climate simula- tion: sensitivity and prediction experiments. *Journal of the Atmospheric Sciences*, 42, 1139–1149.

Mylne, K.R., C. Woolcock, J.C.W. Denholm-Price, and R.J. Darvell, 2002b. Operational calibrated probability forecasts from the ECMWF ensemble prediction system: imple- mentation and verification. Preprints, Symposium on Observations, Data Analysis, and Probabilistic

Prediction, (Orlando, Florida), *American Meteorological Society*, 113–118.

Namias, J., 1952. The annual course of month-to-month persistence in climatic anomalies. *Bulletin of the American Meteorological Society*, 33, 279–285.

National Bureau of Standards, 1959. *Tables of the Bivariate Normal Distribution Function and Related Functions*. Applied Mathematics Series, 50, U.S. Government Printing Office, 258 pp.

Nehrkorn, T., R.N. Hoffman, C. Grassotti, and J.-F. Louis, 2003: Feature calibration and alignment to represent model forecast errors: Empirical regularization. *Quarterly Journal of the Royal Meteorological Society*, 129, 195–218.

Penland, C., and P.D. Sardeshmukh, 1995. The optimal growth of tropical sea surface temperatures anomalies. *Journal of Climate*, 8, 1999–2024.

Peterson, C.R., K.J. Snapper, and A.H. Murphy 1972. Credible interval temperature forecasts. Bulletin of the American Meteorological Society, 53, 966–970.

Pitcher, E.J., 1977. Application of stochastic dynamic prediction to real data. *Journal of the Atmospheric Sciences*, 34, 3–21.

Pitman, E.J.G., 1937. Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society*, B4, 119–130.

Plaut, G., and R. Vautard, 1994. Spells of low-frequency oscillations and weather regimes in the Northern Hemisphere. *Journal of the Atmospheric Sciences*, 51, 210–236.

Tukey, J.W., 1977. *Exploratory Data Analysis*. Reading, Mass., Addison-Wesley, 688 pp.

Unger, D.A., 1985*. A method to estimate the continuous ranked probability score*. Preprints, 9th Conference on Probability and Statistics in the Atmospheric Sciences. American Meteorological Society, 206–213.

Zwiers, F.W., and H.J. Thiébaux, 1987. Statistical considerations for climate experiments. Part I: scalar tests. *Journal of Climate and Applied Meteorology*, 26, 465–476.

Zwiers, F.W., and H. von Storch, 1995. Taking serial correlation into account in tests of the mean. *Journal of Climate*, 8, 336–351.

Lampiran

Tabel A

Kumpulan Contoh Data

Dalam aplikasi nyata dari analisis data klimatologi, kami berharap dapat menggunakan lebih banyak data (misalnya semua data harian Januari yang tersedia, bukan hanya data untuk satu tahun) dan membiarkan komputer melakukan perhitungan. Kumpulan data kecil ini digunakan dalam beberapa contoh di seluruh buku ini sehingga perhitungan dapat dilakukan dengan tangan dan pemahaman yang lebih jelas tentang prosedur dapat dicapai.

Tabel A.1 Pengamatan curah hujan harian (inci) dan suhu (°F) di Ithaca dan Canandaigua, New York, untuk Januari 1987.

Date	Ithaca			Canandaigua		
	Precipitation	Max Temp.	Min Temp.	Precipitation	Max Temp.	Min Temp.
1	0.00	33	19	0.00	34	28
2	0.07	32	25	0.04	36	28
3	1.11	30	22	0.84	30	26
4	0.00	29	−1	0.00	29	19
5	0.00	25	4	0.00	30	16
6	0.00	30	14	0.00	35	24
7	0.00	37	21	0.02	44	26
8	0.04	37	22	0.05	38	24
9	0.02	29	23	0.01	31	24

10	0.05	30	27	0.09	33	29
11	0.34	36	29	0.18	39	29
12	0.06	32	25	0.04	33	27
13	0.18	33	29	0.04	34	31
14	0.02	34	15	0.00	39	26
15	0.02	53	29	0.06	51	38
16	0.00	45	24	0.03	44	23
17	0.00	25	0	0.04	25	13
18	0.00	28	2	0.00	34	14
19	0.00	32	26	0.00	36	28
20	0.45	27	17	0.35	29	19
21	0.00	26	19	0.02	27	19
22	0.00	28	9	0.01	29	17
23	0.70	24	20	0.35	27	22
24	0.00	26	−6	0.08	24	2
25	0.00	9	−13	0.00	11	4
26	0.00	22	−13	0.00	21	5
27	0.00	17	−11	0.00	19	7
28	0.00	26	−4	0.00	26	8
29	0.01	27	−4	0.01	28	14
30	0.03	30	11	0.01	31	14
31	0.05	34	23	0.13	38	23
sum/avg.	3.15	29.87	13	2.4	31.77	20.23
std. dev.	0.243	7.71	13.62	0.168	7.86	8.81

Tabel A.2 Curah hujan Januari di Ithaca, New York, 1933–1982, inci.

1933	0.44	1945	2.74	1958	4.9	1970	1.03
1934	1.18	1946	1.13	1959	2.94	1971	1.11
1935	2.69	1947	2.5	1960	1.75	1972	1.35
1936	2.08	1948	1.72	1961	1.69	1973	1.44
1937	3.66	1949	2.27	1962	1.88	1974	1.84
1938	1.72	1950	2.82	1963	1.31	1975	1.69
1939	2.82	1951	1.98	1964	1.76	1976	3
1940	0.72	1952	2.44	1965	2.17	1977	1.36
1941	1.46	1953	2.53	1966	2.38	1978	6.37
1942	1.3	1954	2	1967	1.16	1979	4.55
1943	1.35	1955	1.12	1968	1.39	1980	0.52
1944	0.54	1956	2.13	1969	1.36	1981	0.87
		1957	1.36			1982	1.51

Tabel A.3 Data Iklim Juni untuk Guayaquil, Ekuador, 1951–1970. Tanda bintang menunjukkan tahun El Niño.

1951∗	26.1	43	1009.5
1952	24.5	10	1010.9
1953∗	24.8	4	1010.7
1954	24.5	0	1011.2
1955	24.1	2	1011.9
1956	24.3	Missing	1011.2
1957∗	26.4	31	1009.3
1958	24.9	0	1011.1
1959	23.7	0	1012
1960	23.5	0	1011.4
1961	24	2	1010.9
1962	24.1	3	1011.5
1963	23.7	0	1011
1964	24.3	4	1011.2
1965∗	26.6	15	1009.9
1966	24.6	2	1012.5
1967	24.8	0	1011.1
1968	24.4	1	1011.8
1969∗	26.8	127	1009.3
1970	25.2	2	1010.6

Tabel B

Tabel Probabilitas

Tabel probabilitas untuk distribusi probabilitas bersama terpilih yang tidak mengikuti ekspresi tertutup untuk fungsi distribusi kumulatif.

Tabel B.1 Probabilitas kumulatif arah kiri untuk distribusi Gaussian standar, $\phi(z) = \Pr\{Z \leq z\}$. Nilai-nilai variabel Gaussian standar z dicantumkan hingga sepersepuluh terdekat di kolom paling kanan dan paling kiri. Judul kolom yang tersisa menunjukkan tempat keseratus mis. Probabilitas sisi kanan diperoleh dengan $\Pr\{Z > z\} = 1 - \Pr\{Z \leq z\}$. Probabilitas untuk $Z > 0$ diperoleh dengan menggunakan simetri distribusi Gaussian, $\Pr\{Z \leq z\} = 1 - \Pr\{Z \leq -z\}$.

Z	0.0900	0.0800	0.0700	0.0600	0.0500	0.0400	0.0300	0.0200	0.0100	0.0000	Z
−4.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	−4.0
−3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	−3.9
−3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	−3.8
−3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	−3.7

−3.6	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	−3.6
−3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	−3.5
−3.4	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
−3.3	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
−3.2	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
−3.1	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
−3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0014	−3.0
−2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
−2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
−2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
−2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0042	0.0043	0.0044	0.0045	0.0047	−2.6
−2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
−2.4	0.0064	0.0066	0.0068	0.0070	0.0071	0.0073	0.0076	0.0078	0.0080	0.0082	−2.4
−2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
−2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0126	0.0129	0.0132	0.0136	0.0139	−2.2
−2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
−2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
−1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
−1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0352	0.0359	−1.8

−1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
−1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
−1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5
−1.4	0.0681	0.0694	0.0708	0.0722	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	−1.4
−1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	−1.3
−1.2	0.0985	0.1003	0.1020	0.1038	0.1057	0.1075	0.1094	0.1112	0.1131	0.1151	−1.2
−1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	−1.1
−1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1563	0.1587	−1.0
−0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	−0.9
−0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	−0.8
−0.7	0.2148	0.2177	0.2207	0.2236	0.2266	0.2297	0.2327	0.2358	0.2389	0.2420	−0.7
−0.6	0.2451	0.2483	0.2514	0.2546	0.2579	0.2611	0.2644	0.2676	0.2709	0.2743	−0.6
−0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	−0.5
−0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	−0.4
−0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	−0.3
−0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4091	0.4129	0.4168	0.4207	−0.2
−0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	−0.1
−0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4841	0.4880	0.4920	0.4960	0.5000	0

Tabel B.2 Kuantil distribusi gamma standar ($\beta = 1$). Item yang ditabulasikan adalah nilai dari variabel acak standar ξ yang sesuai dengan probabilitas kumulatif $F(\xi)$ yang diberikan dalam judul kolom untuk nilai parameter bentuk (α) yang diberikan pada kolom pertama. Untuk menemukan kuantil untuk distribusi dengan parameter skala lainnya, masukkan tabel pada baris yang sesuai, baca nilai standar pada kolom yang sesuai, dan kalikan nilai tabel dengan parameter skala. Untuk mengekstrak probabilitas kumulatif yang sesuai dengan nilai tertentu dari variabel acak, bagi nilai dengan parameter skala, masukkan tabel di baris yang sesuai dengan parameter bentuk, dan interpolasi hasil dari judul kolom

	Cumulative Probability														
a	0.001	0.01	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99	0.999
0.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.007	0.077	0.262	1.057	2.423
0.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.004	0.018	0.070	0.264	0.575	1.554	3.035
0.2	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.006	0.021	0.062	0.164	0.442	0.820	1.894	3.439
0.2	0.000	0.000	0.000	0.000	0.000	0.002	0.007	0.021	0.053	0.122	0.265	0.602	1.024	2.164	3.756
0.3	0.000	0.000	0.000	0.000	0.001	0.006	0.018	0.044	0.095	0.188	0.364	0.747	1.203	2.395	4.024
0.3	0.000	0.000	0.000	0.000	0.003	0.013	0.034	0.073	0.142	0.257	0.461	0.882	1.365	2.599	4.262
0.4	0.000	0.000	0.000	0.001	0.007	0.024	0.055	0.108	0.192	0.328	0.556	1.007	1.515	2.785	4.477
0.4	0.000	0.000	0.000	0.002	0.013	0.038	0.080	0.145	0.245	0.398	0.644	1.126	1.654	2.958	4.677

0.5	0.000	0.000	0.001	0.005	0.022	0.055	0.107	0.186	0.300	0.468	0.733	1.240	1.786	3.121	4.863
0.5	0.000	0.000	0.002	0.008	0.032	0.074	0.138	0.228	0.355	0.538	0.819	1.349	1.913	1.124	5.040
0.6	0.000	0.000	0.004	0.012	0.045	0.096	0.170	0.272	0.411	0.607	0.904	1.454	2.034	3.421	5.208
0.6	0.000	0.000	0.006	0.018	0.059	0.120	0.204	0.316	0.467	0.676	0.987	1.556	2.150	3.562	5.370
0.7	0.000	0.001	0.009	0.025	0.075	0.146	0.240	0.362	0.523	0.744	1.068	1.656	2.264	3.698	5.526
0.7	0.000	0.001	0.012	0.033	0.093	0.173	0.276	0.408	0.579	0.811	1.149	1.753	2.374	3.830	5.676
0.8	0.000	0.002	0.017	0.043	0.112	0.201	0.314	0.455	0.636	0.878	1.227	1.848	2.481	3.958	5.822
0.8	0.000	0.003	0.022	0.053	0.132	0.231	0.352	0.502	0.692	0.945	1.305	1.941	2.586	4.083	5.964
0.9	0.000	0.004	0.028	0.065	0.153	0.261	0.391	0.550	0.749	1.010	1.382	2.032	2.689	4.205	6.103
0.9	0.001	0.006	0.035	0.078	0.176	0.292	0.431	0.598	0.805	1.076	1.458	2.122	2.790	4.325	6.239
1	0.001	0.008	0.043	0.091	0.199	0.324	0.471	0.646	0.861	1.141	1.533	2.211	2.888	4.441	6.373
1	0.001	0.011	0.052	0.106	0.224	0.357	0.512	0.694	0.918	1.206	1.607	2.298	2.986	4.556	6.503
1.1	0.002	0.013	0.061	0.121	0.249	0.391	0.553	0.742	0.974	1.270	1.681	2.384	3.082	4.669	6.631
1.1	0.002	0.017	0.071	0.138	0.275	0.425	0.594	0.791	1.030	1.334	1.759	2.469	1.27	4.781	6.757
1.2	0.002	0.020	0.082	0.155	0.301	0.459	0.636	0.840	1.086	1.397	1.831	2.553	1.120	4.890	6.881
1.2	0.002	0.024	0.094	0.173	0.329	0.494	0.678	0.889	1.141	1.460	1.903	2.636	1.212	4.998	7.003
1.3	0.003	0.027	0.106	0.191	0.357	0.530	0.720	0.938	1.197	1.523	1.974	2.719	3.453	5.105	7.124
1.3	0.004	0.032	0.119	0.210	0.385	0.566	0.763	0.987	1.253	1.586	2.045	2.800	3.544	5.211	7.242
1.4	0.004	0.037	0.133	0.230	0.414	0.602	0.806	1.036	1.308	1.649	2.115	2.881	3.633	5.314	7.360
1.4	0.005	0.043	0.145	0.250	0.443	0.639	0.849	1.085	1.364	1.711	2.185	2.961	3.722	5.418	7.476
1.5	0.007	0.049	0.160	0.272	0.473	0.676	0.892	1.135	1.419	1.773	2.255	3.041	3.809	5.519	7.590

1.5	0.008	0.056	0.175	0.293	0.504	0.713	0.935	1.184	1.474	1.834	2.324	3.120	3.897	5.620	7.704
1.6	0.011	0.063	0.191	0.313	0.534	0.750	0.979	1.234	1.530	1.896	2.392	1.49	3.983	5.720	7.816
1.6	0.014	0.071	0.207	0.336	0.565	0.788	1.023	1.283	1.585	1.957	2.461	1.126	4.068	5.818	7.928
1.7	0.018	0.078	0.224	0.359	0.597	0.826	1.067	1.333	1.640	2.018	2.529	1.204	4.153	5.917	8.038
1.7	0.023	0.087	0.241	0.382	0.628	0.865	1.111	1.382	1.695	2.079	2.597	3.431	4.237	6.014	8.147
1.8	0.031	0.096	0.259	0.406	0.661	0.903	1.155	1.432	1.750	2.140	2.664	3.507	4.321	6.110	8.255
1.8	0.036	0.104	0.277	0.430	0.693	0.942	1.199	1.481	1.805	2.200	2.731	3.584	4.405	6.207	8.362
1.9	0.041	0.115	0.296	0.454	0.726	0.980	1.244	1.531	1.860	2.261	2.798	3.659	4.487	6.301	8.469
1.9	0.045	0.124	0.314	0.479	0.759	1.020	1.288	1.580	1.915	2.321	2.865	3.735	4.569	6.396	8.575
2	0.049	0.136	0.334	0.505	0.790	1.059	1.333	1.630	1.969	2.381	2.931	3.809	4.651	6.490	8.679
2	0.053	0.151	0.354	0.530	0.823	1.099	1.378	1.680	2.024	2.442	2.997	3.883	4.732	6.582	8.783
2.1	0.057	0.164	0.374	0.556	0.857	1.138	1.422	1.729	2.079	2.501	3.063	3.958	4.813	6.675	8.887
2.1	0.066	0.175	0.395	0.583	0.891	1.178	1.467	1.779	2.133	2.561	3.129	4.032	4.894	6.767	8.989
2.2	0.070	0.186	0.415	0.610	0.925	1.218	1.512	1.829	2.188	2.620	1.45	4.105	4.973	6.858	9.091
2.2	0.074	0.200	0.437	0.637	0.959	1.258	1.557	1.879	2.242	2.680	1.110	4.179	5.053	6.949	9.193
2.3	0.085	0.212	0.458	0.664	0.994	1.298	1.603	1.928	2.297	2.739	1.175	4.252	5.132	7.039	9.294
2.3	0.090	0.226	0.481	0.691	1.029	1.338	1.648	1.978	2.351	2.799	1.240	4.324	5.211	7.129	9.394
2.4	0.095	0.238	0.502	0.718	1.064	1.379	1.693	2.028	2.405	2.858	3.455	4.396	5.289	7.219	9.493
2.4	0.100	0.253	0.524	0.747	1.099	1.420	1.738	2.078	2.459	2.917	3.519	4.468	5.367	7.308	9.592

2.5	0.113	0.268	0.548	0.775	1.134	1.460	1.784	2.127	2.514	2.976	3.584	4.540	5.445	7.397	9.691
2.5	0.118	0.280	0.575	0.804	1.170	1.500	1.829	2.178	2.568	3.035	3.648	4.612	5.522	7.484	9.789
2.6	0.124	0.296	0.598	0.833	1.205	1.539	1.875	2.227	2.622	3.093	3.712	4.683	5.600	7.572	9.886
2.6	0.130	0.313	0.621	0.862	1.241	1.581	1.920	2.277	2.676	3.152	3.776	4.754	5.677	7.660	9.983
2.7	0.147	0.326	0.646	0.890	1.277	1.622	1.966	2.327	2.730	1.60	3.840	4.825	5.753	7.746	10.079
2.7	0.152	0.343	0.671	0.920	1.314	1.663	2.011	2.376	2.784	1.119	3.903	4.896	5.830	7.833	10.176
2.8	0.158	0.356	0.694	0.950	1.350	1.704	2.058	2.427	2.838	1.178	3.967	4.966	5.906	7.919	10.272
2.8	0.165	0.374	0.719	0.980	1.386	1.746	2.103	2.476	2.892	1.236	4.030	5.040	5.982	8.004	10.367
2.9	0.186	0.392	0.744	1.009	1.423	1.787	2.149	2.526	2.946	3.444	4.093	5.120	6.058	8.090	10.461
2.9	0.192	0.406	0.770	1.040	1.460	1.829	2.195	2.576	2.999	3.502	4.156	5.190	6.133	8.175	10.556
3	0.198	0.424	0.794	1.070	1.497	1.871	2.241	2.626	3.054	3.560	4.220	5.260	6.208	8.260	10.649
3	0.205	0.439	0.819	1.101	1.534	1.913	2.287	2.676	3.108	3.618	4.283	5.329	6.283	8.345	10.743
3.1	0.212	0.458	0.845	1.134	1.571	1.954	2.333	2.726	1.11	3.676	4.346	5.398	6.357	8.429	10.837
3.1	0.239	0.478	0.872	1.165	1.607	1.996	2.378	2.776	1.65	3.734	4.408	5.468	6.432	8.513	10.930
3.2	0.245	0.492	0.898	1.197	1.645	2.038	2.425	2.825	1.118	3.792	4.471	5.537	6.506	8.596	11.023
3.2	0.251	0.513	0.925	1.227	1.682	2.080	2.471	2.875	1.172	3.850	4.533	5.605	6.580	8.680	11.113
3.3	0.259	0.528	0.950	1.259	1.720	2.123	2.517	2.925	1.226	3.907	4.595	5.675	6.654	8.763	11.205
3.3	0.267	0.548	0.977	1.291	1.758	2.165	2.563	2.975	3.430	3.965	4.658	5.743	6.727	8.845	11.298
3.4	0.300	0.570	1.004	1.323	1.796	2.207	2.610	3.025	3.483	4.022	4.720	5.811	6.801	8.928	11.389
3.4	0.306	0.585	1.031	1.354	1.834	2.250	2.656	3.075	3.537	4.079	4.782	5.879	6.874	9.010	11.480
3.5	0.313	0.607	1.059	1.386	1.872	2.292	2.702	3.125	3.590	4.137	4.843	5.948	6.947	9.093	11.570

3.5	0.320	0.623	1.087	1.418	1.910	2.334	2.748	1.25	3.644	4.194	4.905	6.015	7.020	9.174	11.660
3.6	0.328	0.645	1.115	1.451	1.948	2.377	2.795	1.75	3.697	4.252	4.967	6.084	7.092	9.255	11.749
3.6	0.337	0.661	1.141	1.483	1.985	2.420	2.841	1.124	3.750	4.309	5.028	6.152	7.165	9.337	11.840
3.7	0.377	0.684	1.169	1.516	2.024	2.462	2.887	1.174	3.804	4.366	5.091	6.219	7.237	9.418	11.929
3.7	0.383	0.708	1.197	1.549	2.062	2.505	2.934	1.224	3.858	4.423	5.152	6.286	7.310	9.499	12.017
3.8	0.390	0.723	1.226	1.582	2.101	2.547	2.980	3.425	3.911	4.480	5.214	6.354	7.381	9.579	12.107
3.8	0.398	0.748	1.255	1.613	2.140	2.590	3.027	3.474	3.964	4.537	5.275	6.420	7.454	9.659	12.195
3.9	0.406	0.764	1.284	1.646	2.179	2.633	3.073	3.524	4.018	4.594	5.336	6.488	7.525	9.740	12.284
3.9	0.416	0.788	1.310	1.680	2.218	2.676	3.120	3.574	4.071	4.651	5.397	6.555	7.596	9.820	12.371
4	0.426	0.805	1.339	1.713	2.257	2.719	1.13	3.624	4.124	4.708	5.458	6.622	7.668	9.900	12.459
4	0.471	0.829	1.369	1.746	2.295	2.762	1.59	3.674	4.177	4.765	5.519	6.689	7.739	9.980	12.546
4.1	0.478	0.847	1.398	1.780	2.334	2.805	1.106	3.724	4.231	4.822	5.580	6.755	7.811	10.059	12.634
4.1	0.485	0.871	1.429	1.814	2.373	2.848	1.152	3.774	4.284	4.879	5.641	6.821	7.882	10.137	12.721
4.2	0.494	0.900	1.455	1.848	2.413	2.891	1.200	3.823	4.337	4.936	5.701	6.888	7.952	10.216	12.807
4.2	0.503	0.914	1.485	1.882	2.451	2.935	1.246	3.874	4.390	4.992	5.762	6.954	8.023	10.295	12.894
4.3	0.513	0.942	1.515	1.916	2.491	2.978	3.443	3.924	4.444	5.049	5.823	7.020	8.093	10.374	12.981
4.3	0.524	0.958	1.545	1.950	2.531	3.021	3.489	3.974	4.497	5.105	5.883	7.086	8.170	10.453	13.066
4.4	0.578	0.986	1.576	1.985	2.572	3.065	3.537	4.024	4.550	5.162	5.944	7.153	8.264	10.531	13.152
4.4	0.584	1.002	1.603	2.017	2.612	3.108	3.584	4.074	4.603	5.218	6.005	7.219	8.334	10.609	11.88

4.5	0.592	1.029	1.634	2.051	2.653	3.152	3.630	4.123	4.656	5.274	6.065	7.284	8.405	10.687	11.174
4.5	0.600	1.046	1.665	2.085	2.691	1.45	3.677	4.173	4.709	5.331	6.126	7.350	8.475	10.765	11.260
4.6	0.610	1.074	1.696	2.120	2.731	1.89	3.724	4.223	4.762	5.387	6.186	7.415	8.544	10.843	13.495
4.6	0.620	1.092	1.727	2.155	2.771	1.133	3.771	4.273	4.815	5.443	6.246	7.480	8.615	10.920	13.578
4.7	0.632	1.119	1.755	2.190	2.812	1.176	3.817	4.323	4.868	5.501	6.306	7.546	8.684	10.998	13.663
4.7	0.698	1.138	1.786	2.225	2.852	1.219	3.864	4.373	4.921	5.557	6.366	7.611	8.754	11.075	13.748
4.8	0.703	1.165	1.817	2.260	2.890	1.262	3.911	4.423	4.974	5.613	6.426	7.676	8.823	11.152	13.832
4.8	0.710	1.184	1.849	2.295	2.930	3.456	3.958	4.474	5.027	5.669	6.486	7.742	8.892	11.229	13.916
4.9	0.717	1.211	1.881	2.330	2.970	3.500	4.005	4.524	5.081	5.725	6.546	7.807	8.962	11.306	14.000
4.9	0.726	1.247	1.909	2.366	3.011	3.544	4.052	4.573	5.134	5.781	6.606	7.872	9.031	11.382	14.084
5	0.737	1.258	1.940	2.398	3.051	3.588	4.099	4.623	5.186	5.837	6.665	7.937	9.100	11.457	14.168
5	0.748	1.293	1.972	2.434	3.091	3.632	4.146	4.673	5.239	5.893	6.725	8.002	9.169	11.534	14.251

Tabel B.3 Kuantil kanan dari distribusi Chi-kuadrat. Untuk v besar, distribusi chi-kuadrat kira-kira Gaussian, dengan rata-rata v dan varians 2v.

	Cumulative Probability					
V	0.5	0.9	0.95	0.99	0.999	0.9999
1	0.455	2.706	3.841	6.635	10.828	15.137
2	1.386	4.605	5.991	9.21	13.816	18.421

3	2.366	6.251	7.815	11.345	16.266	21.108
4	1.207	7.779	9.488	11.127	18.467	23.512
5	4.351	9.236	11.07	15.086	20.515	25.745
6	5.348	10.645	12.592	16.812	22.458	27.855
7	6.346	12.017	14.067	18.475	24.322	29.878
8	7.344	11.212	15.507	20.09	26.124	31.827
9	8.343	14.684	16.919	21.666	27.877	33.719
10	9.342	15.987	18.307	21.59	29.588	35.563
11	10.341	17.275	19.675	24.725	31.264	37.366
12	11.34	18.549	21.026	26.217	32.91	39.134
13	12.34	19.812	22.362	27.688	34.528	40.871
14	11.189	21.064	23.685	29.141	36.123	42.578
15	14.339	22.307	24.996	30.578	37.697	44.262
16	15.338	23.542	26.296	32	39.252	45.925
17	16.338	24.769	27.587	31.259	40.79	47.566
18	17.338	25.989	28.869	34.805	42.312	49.19
19	18.338	27.204	30.144	36.191	43.82	50.794
20	19.337	28.412	31.41	37.566	45.315	52.385
21	20.337	29.615	32.671	38.932	46.797	53.961

22	21.337	30.813	33.924	40.289	48.268	55.523
23	22.337	32.007	35.172	41.638	49.728	57.074
24	21.187	31.46	36.415	42.98	51.179	58.613
25	24.337	34.382	37.652	44.314	52.62	60.14
26	25.336	35.563	38.885	45.642	54.052	61.656
27	26.336	36.741	40.113	46.963	55.476	61.14
28	27.336	37.916	41.337	48.278	56.892	64.661
29	28.336	39.087	42.557	49.588	58.301	66.152
30	29.336	40.256	43.773	50.892	59.703	67.632
31	30.336	41.422	44.985	52.191	61.098	69.104
32	31.336	42.585	46.194	53.486	62.487	70.57
33	32.336	43.745	47.4	54.776	63.87	72.03
34	31.186	44.903	48.602	56.061	65.247	73.481
35	34.336	46.059	49.802	57.342	66.619	74.926
36	35.336	47.212	50.998	58.619	67.985	76.365
37	36.336	48.363	52.192	59.892	69.347	77.798
38	37.335	49.513	51.234	61.162	70.703	79.224
39	38.335	50.66	54.572	62.428	72.055	80.645
40	39.335	51.805	55.758	63.691	71.252	82.061
41	40.335	52.949	56.942	64.95	74.745	83.474
42	41.335	54.09	58.124	66.206	76.084	84.88

43	42.335	55.23	59.304	67.459	77.419	86.28
44	41.185	56.369	60.481	68.71	78.75	87.678
45	44.335	57.505	61.656	69.957	80.077	89.07
46	45.335	58.641	62.83	71.201	81.4	90.456
47	46.335	59.774	64.001	72.443	82.721	91.842
48	47.335	60.907	65.171	73.683	84.037	91.71
49	48.335	62.038	66.339	74.919	85.351	94.597
50	49.335	61.17	67.505	76.154	86.661	95.968
55	54.335	68.796	71.161	82.292	91.18	102.776
60	59.335	74.397	79.082	88.379	99.607	109.501
65	64.335	79.973	84.821	94.422	105.988	116.16
70	69.334	85.527	90.531	100.425	112.317	122.754
75	74.334	91.061	96.217	106.393	118.599	129.294
80	79.334	96.578	101.879	112.329	124.839	135.783
85	84.334	102.079	107.522	118.236	131.041	142.226
90	89.334	107.565	113.145	124.116	137.208	148.626
95	94.334	113.038	118.752	129.973	141.194	154.989
100	99.334	118.498	124.342	135.807	149.449	161.318