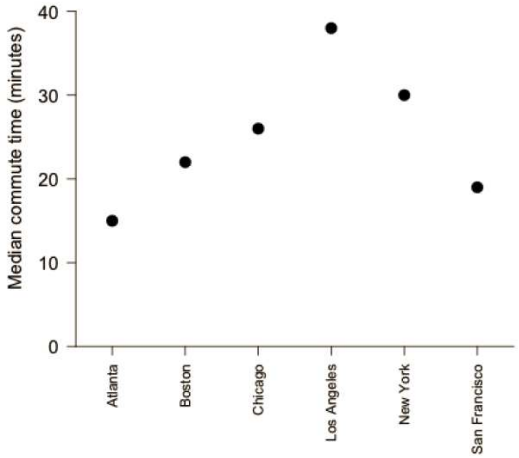
5 Bab 5 Visualisasi Data atau Visualisasi Statistik?

Bab ini membahas kapan menggunakan visualisasi data dan visualisasi ststistik.

Setiap dilakukan peringkasan data maka disitulah akan digunakan statistik. *Mean* merupakan nilai tengah data yang diperoleh menjumlahkan semua data membaginya dengan jumlah data. *Median* merupakan nilai yang memiliki setengah data berada di atas (atau sama dengan) nilai tersebut dan setengahnya data berada di bawah (atau sama dengan) nilai tersebut. Persentase data yang termasuk dalam suatu kategori juga merupakan statistik. Secara teknis, bahkan ketinggian dari bilah histogram juga merupakan statistik.

Ada pilihan antara menampilkan data individual dan statistik, dan tujuannya harus jelas. Memang seluruh data dapat ditampilkan sekaligus, tetapi hal ini juga mungkin membuat pembaca sulit untuk menangkap pesan. Maka mungkin akan lebih tepat sasaran jika hanya statistik data saja yang ditampilkan.

Sebagai contoh misalnya ingin dibandingkan lama perjalanan orang dari rumah ke tempat kerja pada beberapa kota di Amerika, maka statistik ringkasan untuk masing-masing kota (Gambar 5.1) sudah cukup membantu. Jika yang ditunjukkan hanya penanda untuk setiap statistik, ini disebut *dot plot*.



Gambar 5.1 *Dot plot* yang menunjukkan rata-rata (atau statistik lain) untuk grup-grup dalam data, atau variabel. *Dot plot* bisa juga memiliki statistik yang dikodekan dalam posisi horizontal.

.

5.1 Memilih data atau statistik

Terlepas dari tujuan kejelasan, visualisasi harus mempertimbangkan kebutuhan pembaca. Jika pembaca hanya perlu untuk mengetahui statistik, misalnya median harga rumah di suatu lingkungan, maka jangan buang waktu dengan menampilkan harga setiap rumah. Ada perangkat visualisasi interaktif online tersedia yang memungkinkan fleksibilitas dalam hal ini. Pembuatan visualisasi data perlu memperhatikan batas-batas perjanjian atau hukum yang mengaturnya. Ada keadaan dimana sumber data yang berpartisipasi dalam survey mungkin menolak untuk dipublikasi.

Visualisasi hanya ringkasan statistik dapat dilakukan sepanjang dapat mencapai sasarannya. Tapi menggabungkan statistik dengan data, juga dapat dilakukan sepanjang visualisasinya tidak berantakan. Para ahli menyarankan penggunakan *spike histogram* (histogram dengan sedikit puncak memanjang turun dari sumbu horizontal ke posisi *means*), kuartil atau lainnya untuk visualisasi. Visualisasi seperti ini kompak dan memungkinkan pembaca untuk memilih apa yang harus dilihat.

Semua statistik yang relevan harus ditampilkan dan bukan hanya statistik yang terlihat menarik saja. Ada dilema yang mungkin

dihadapi, menonjolkan statistik tertentu mungkin berakibat menyembunyikan fakta sebernarnya.

5.2 Standar Deviasi

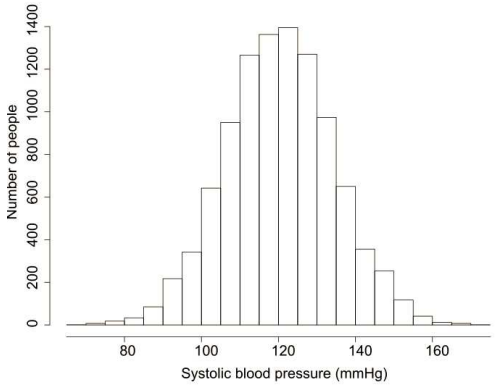
Mean (rata-rata) adalah suatu bilangan yang mewakili sekumpulan data. Dalam statistika, **rata-rata**, **rerata** atau *means* memiliki tiga arti yang berkaitan:

- rerata aritmetik, pengertian yang paling umum dikenal awam,
- nilai harapan dari suatu peubah acak, dan
- ukuran pemusatan dari suatu sebaran probabilitas.

Rerata merupakan salah satu konsep sentral dalam statistika matematis, dan bersama dengan *varians* menjadi bagian penting dalam berbagai penurunan berbagai metode statistika.

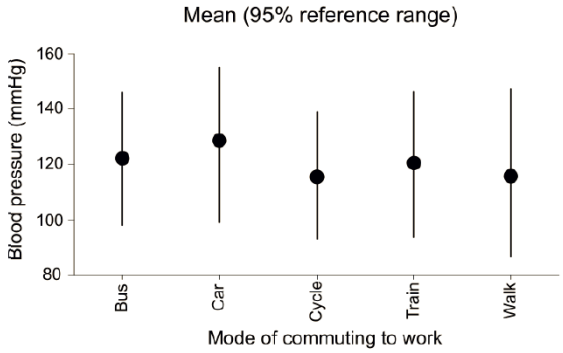
Dipandang dari sisi matematis, rerata adalah momen pertama dari suatu peubah acak. Momen pertama mengenai rerata dari suatu peubah acak disebut *simpangan* (deviasi).

Distribusi normal di Gambar 5.2 sering ditemui dalam variabel kontinu, dan jika data memiliki histogram seperti ini, rerata akan berada di titik tertinggi, dan standar deviasi adalah jarak horizontal dari *mean* ke titik di mana bilah histogram berubah menggembung ke atas hingga menggembung ke bawah.



Gambar 5.2 Histrogram dari suatu distribusi normal.

Cara yang lebih intuitif untuk mengungkapkan distribusi ini dalam kata-kata adalah bahwa 95% dari data harus terletak di antara dua standar deviasi di bawah *mean* dan dua standar deviasi di atas *mean*. Ini berlaku untuk data yang didistribusikan secara normal, bukan dengan bentuk lain dan kadang-kadang disebut rentang referensi 95%. Visualisasi *error bar* memanjang di atas dan di bawah rata-rata ditunjukkan dalam Gambar 5.3.



Gambar 5.3 *Dot plot* dengan *error bar*.

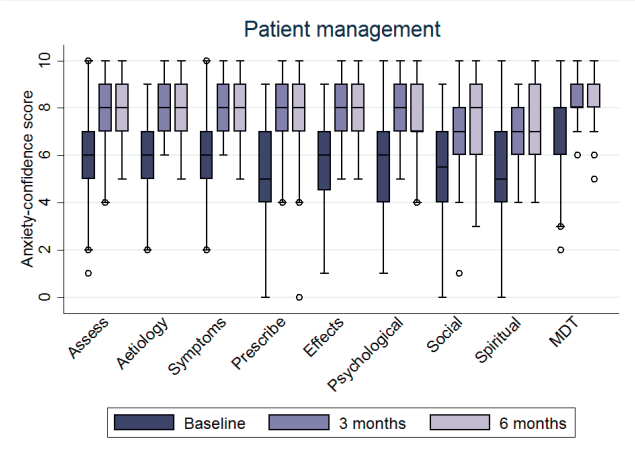
Dalam dua dimensi (seperti pada *scatter plot*) rentang referensi untuk variabel horizontal, dan yang lainnya untuk vertikal dapat dibuat, dan divisualisasikannya sebagai dua set *error bar* di sudut kanan, atau sebagai elips.

5.3 Kuartil dan statistik lainnya

Median memiliki interpretasi yang lebih cepat daripada *mean*. Setengah dari data terletak pada median atau di atasnya, sisanya terletak pada median atau di bawahnya. Ini juga lebih *robust*, jika ada kesalahan dalam pengumpulan data yang muncul sebagai *outlier* (pencilan) yang sangat rendah atau sangat tinggi, *outlier* akan menarik *mean* ke atas dan ke bawah. Sebaliknya karena median tidak terlalu terpengaruh oleh *outlier* karena memiliki memiliki setengah data di atas dan setengah di bawah. Tidak masalah apakah pencilan jauh di atas atau di bawah, median tidak terpengaruh.
Karena itu *median* adalah alternatif yang kuat untuk *mean*. Selain itu ada statistik lain yang dapat digunakan untuk menggambarkan penyebaran data. Yang paling umum adalah *kuartil*, yang membelah data menjadi empat bagian yang sama. Seperempat data berada di

bawah (atau pada) kuartil pertama, dan seperempat berada pada kuartil ketiga. Kuartil kedua sama dengan median. Kuartil pertama juga dikenal sebagai *centile* ke-25 (bayangkan data dibagi menjadi 100 bagian), *median* adalah *centile* ke-50, dan kuartil ketiga adalah *centile* ke-75.

Dot plot dapat digunakan untuk menempatkan *marker* di *median* dan *error bar* memanjang ke beberapa kuantil, selama dibuat dengan jelas apa yang mereka wakili. Format yang populer, memperluas ide ini adalah *box plot* (Gambar 5.4).



Gambar 5.4 *Box plot* yang membandingkan sembilan variabel pada tiga waktu.

Kata *quantile*/kuantil berasal dari kata *quantity*. Secara sederhana, kuantil adalah tempat dimana sampel dibagi menjadi subkelompok yang berukuran sama. *Median* adalah kuantil; median ditempatkan dalam distribusi probabilitas sehingga tepat setengah dari data lebih rendah dari median dan setengah dari data di atas median. Median memotong distribusi menjadi dua bidang yang sama sehingga terkadang disebut 2-kuantil.

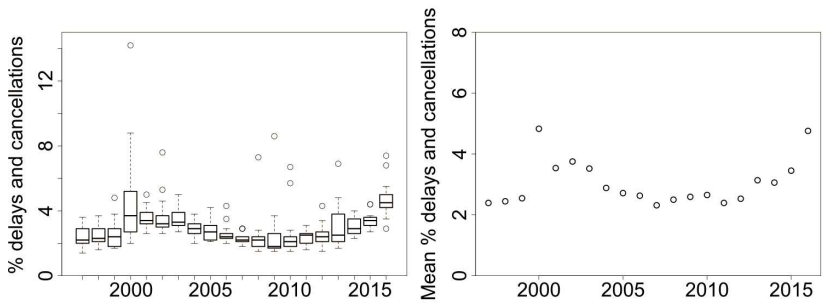
Kuartil juga merupakan kuantil; kuartil membagi distribusi menjadi empat bagian yang sama. *Persentil* adalah kuantil yang membagi distribusi menjadi 100 bagian yang sama dan *desil* adalah kuantil yang membagi distribusi menjadi 10 bagian yang sama.

Median dapat juga disebut sebagai 0,5 kuantil, yang berarti bahwa proporsi 0,5 (setengah) bearada di bawah median dan 0,5 berada di atas median.

Jika sebuah *box plot* memiliki lima nilai statistik, dan disandikan ke posisi horizontal atau posisi vertikal, maka m*edian* adalah garis di tengah kotak, dan tepi kotak adalah kuartil pertama dan ketiga.

Selain *quantile* statistik yang tangguh adalh melalui pemangkasan data (*data trimimng*). Pemangkasan data dilakukan dengan menyisihkan data di luar jumlah tertentu sebelum menghitung rata-rata atau nilai statistik lainnya. Sebagai contoh memangkas 10% data teratas dan 10% data terbawah akan menghasilkan rata-rata dari 80% data. Teknik lain yaitu *Winsorizing* dilakukan dengan mengganti nilai data luar dengan yang kuantil yang dipilih sebagai ambang, sehingga secara efektif data luar tesebut dengan kuantil, dan kemudian statistik dihitung. Pada kasus 80% data sisa pemangkasan atau hasil *Winsorized* median yang dihasilkan tidak akan berbeda dengan median asli. Karena itu rentang antar-kuartil tidak akan berubah sampai data dipotong atau *Winsorize* sebesar 50% atau lebih.

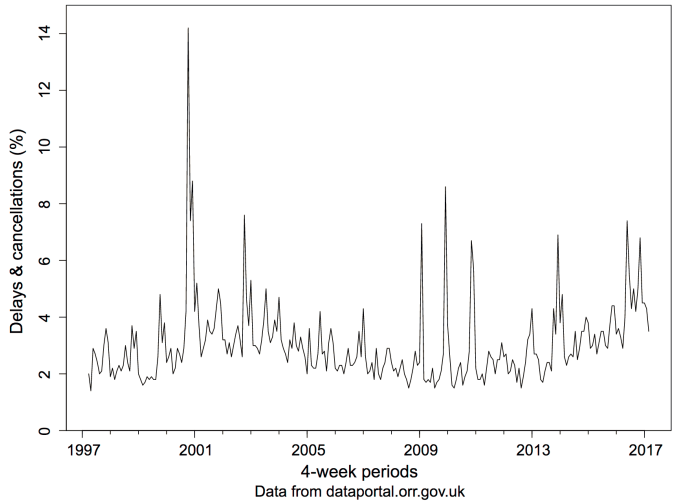
Untuk menggambarkan sebaran data, standar deviasi dan kisaran antar kuartil dapat dipilih, dan deviasi absolut rata-rata juga dapat digunakan. Nilai rata-rata dapat dihitung dan kemudian nilai jarak masing-masing (deviasi) di atas atau di bawah nilai rata-rata ditetapkan. Deviasi negatif dapat diubah ke positif, sehingga median dapat dicari. Cara ini adalah cara yang baik untuk menggambarkan pencaran disekitar *mean*.



Gambar 5.5 *Box plot* (kiri) dan *dot plot* (kanan) menggambarkan perubahan *delay* perjalanan kereta selama 20 tahun. Perhatikan *mean* pada tahun 2000 meningkat pesat yang patut dicurigai sebagai oulier di *box plot*.

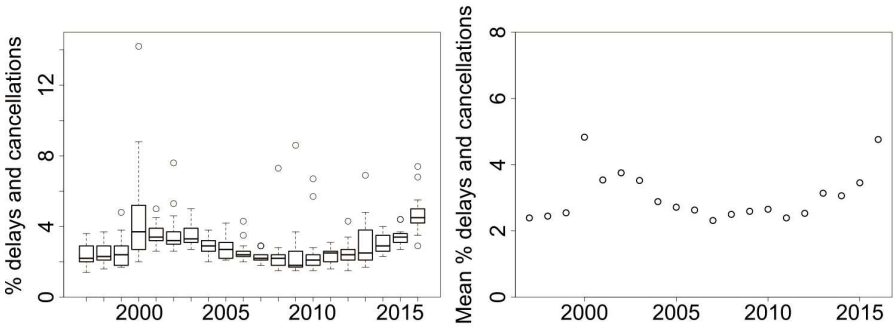
5.4 *Smoothing*

Pembahasan sebelumnya telah menjelaskan beberapa statistik seperti mean, standard deviation, quantile dan lain-lain. Namun kadangkala visualisasi membutuhkan lebih dari itu. Perhatikan kasus sebelumnya (Gambar 2.2) kurang dapat menunjukkan trend dan pola karena terlalu banyak data yang ditampilkan. Untuk dapat menunjukkan trend dan pola yang lebih umum dibutuhkan pengurangan jumlah data.



Gambar 2.2 Grafik garis keterlambatan kereta.

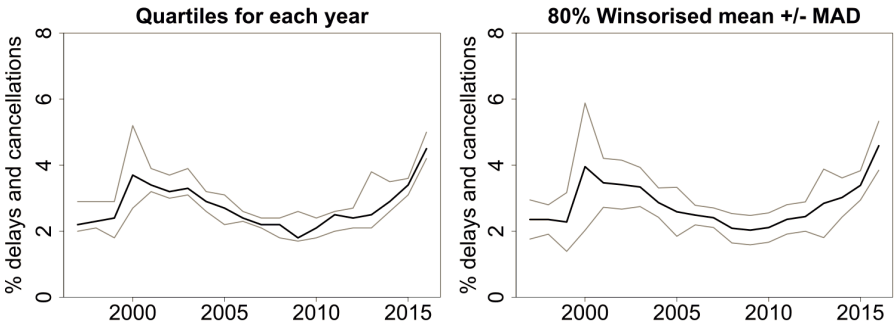
Untuk memvisualisasikan tren jangka panjang, waktu dapat disederhanakan/dipotong menjadi tahun dan kemudian digambar beberapa statistik untuk setiap tahun. Gambar 5.5 menunjukkan *box plot* dan *dot plot*, sedangkan Gambar 5.6 menggabungkan beberapa nilai statistik dari waktu ke waktu: *kuartil* dan 80% *Winsorized mean* ± deviasi sekitar median. Karena ada 13 pengamatan di setiap tahun, *Winsorisasi* 80% mempengaruhi dua periode terburuk dan dua periode terbaik.



Gambar 5.5 Sebuah *box plot* (kiri) dan *dot plot* (kanan) meringkas perubahan keterlambatan kereta selama 20 tahun. Perhatikan *mean* untuk untuk tahun 2000 meningkat drastis dan dapat diduga sebagai pencilan dalam *box plot*.

Dari dua pendekatan di atas, tidak satu pun dari opsi ini yang benar atau salah. Pilihan terbaik adalah opsi mana yang paling membantu pembaca untuk paling memahami pesandengan mudah. Dapat dikatakan opsi untuk meringkas data dan memvisualisasikan ringkasan selalu banyak alternatif.

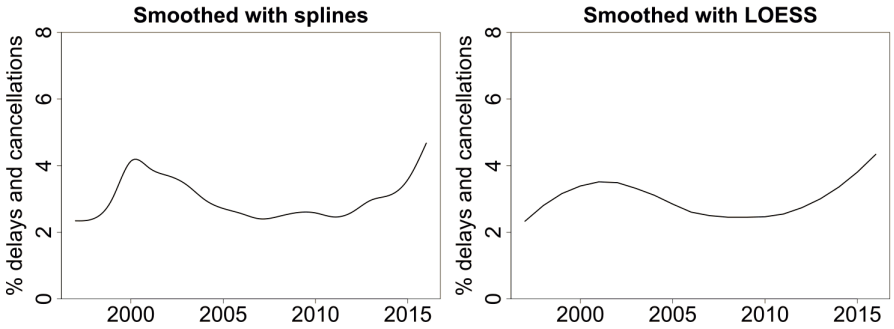
Ada beberapa teknik untuk memuluskan grafik garis. Grafik yang mulus akan memudahkan pembaca untuk menyerap mencerna informasi, karena informasi yang ada dalam visualisasi menjadi tidak rumit.



Gambar 5.6 Grafik Garis yang menghubungkan statistik yang merangkum perubahan pada data keterlambatan kereta selama 20 tahun: kuartil (kiri) dan 80% *Winsorized mean* ± deviasi absolut (kanan).

Telah dibahas perbandingan *kernel density plot* dengan histogram pada bab 3. Beberapa Metode paling populer untuk menempatkan garis halus melalui *scatter plot* disebut *splines*, LOESS dan *polished median* dan ini tersedia pada perangkat lunak analisis data.

Dalam visualisasi dapat diputuskan seberapa banyak *smoothing* diterapkan. Jika terlalu banyak smoothing maka informasi detail penting bisa hilang, namun jika terlalu sedikit dan grafik garis yang terbentuk tetap terlihat bergelombang dan mungkin mengganggu pembaca untuk melihat tren yang penting.



Gambar 5.7 Data yang diperhalus (*Smoothing*) untuk meringkas data perubahan dalam keterlambatan kereta selama 20 tahun menggunakan *splines* (kiri) dan LOESS (kanan).