4 Bab 4: Persentase dan Risiko

Bab ini membahas data berbentuk Persentase dan Risiko.

Saat data hanya terdiri dari "ya" dan "tidak" untuk setiap pengamatan, maka representasi statisitiknya sangat terbatas. Tapi bukan berarti bahwa data seperti ini tidak berguna. Data seperti ini adalah keanggotaan dari suatu kategori, dan “ya” dan “tidak” merupakan kategori yang paling sederhana. Data yang diisi responden dalam survei dengan cara mencentang kotak, akan menempatkan pilihan ke dalam kategori. Mungkin hanya ada dua opsi: dicentang atau tidak dicentang, atau ada beberapa kategori.

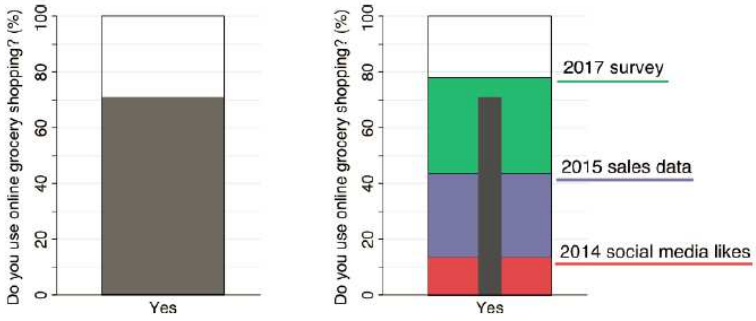
4.1 Visualisasi Satu variabel pada Satu waktu

Untuk menggambarkan variabel dengan dua kategori, seperti "ya" dan "tidak," hanya tiga angka yang dibutuhkan yaitu : berapa banyak yang mengatakan "ya," berapa banyak yang menjawab pertanyaan, dan berapa persen itu. Jika dari survei 200 orang dan didapat 60 mengatakan "ya," berarti 60/200 = 30%. Kadang orang menggunakan proporsi diamana membagi angka yang mengatakan "ya" dengan angka yang menjawab, jadi 30%, ditulis sebagai proporsi adalah 0,3. Dua cara ini secara visual akan terlihat sama kecuali pada label pada sumbunya. Namun sebagian besar pembaca lebih terbiasa melihat persentase. Bagaimana dengan data orang yang mengatakan "tidak"? Secara implisit berarti 200 - 60 = 140, jadi tidak perlu dituliskan. Tapi jika ada beberapa jawaban "ya", beberapa "tidak," dan beberapa kosong, maka data kosong harus diperhitungkan. Dalam hal ini dapat dibagi menjadi tiga kategori (“60 (30%) mengatakan ya, 110 (55%) mengatakan tidak, dan 30

(15%) tidak menjawab ”). Atau dapat disebutkan berapa banyak yang kosong lalu kurangi total dengan jawaban "ya" dan “Tidak”: 60/180 = 33% mengatakan ya, sedangkan 20 tidak menjawab pertanyaan. Persentase sering digambar dalam visualisasi. Jadi harus dipastikan penghitungan setiap kategori tidak hilang dan sertakan dalam label atau teks yang menyertainya.

Persentase sederhana paling baik disandikan dalam bentuk panjang, misalnya grafik batang/*bar chart* atau plot titik/*dot plot*. Agar sedikit lebih menarik, bisa digunakan *pictogram* (lihat Gambar 2.8), yang juga disandikan menjadi bentuk panjang. Untuk memberi kesan keseluruhan sebagai 100%, panjangnya dapat ditunjukkan dengan latar belakang 100% (Gambar 4.1 kiri), atau untuk membandingkan dengan beberapa target atau nilai referensi lainnya, variasi tentang ini disebut *bullet chart* (Gambar 4.1 kanan).

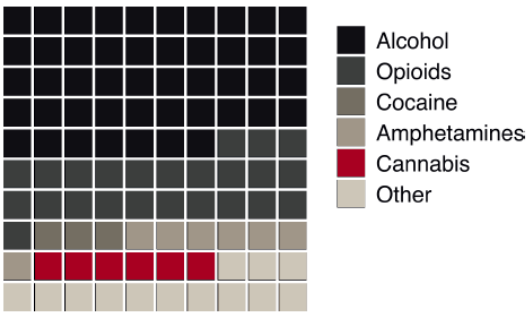
Ketika ada lebih dari dua kategori, semua kategori harus terlihat atau tercantum dalam teks yang menyertainya, meskipun hanya untuk menunjukkan orang yang abstain. Misalnya, pelanggan yang membeli barang secara online diklasifikasikan sesuai dengan negara asal mereka (di mana alamat penagihan adalah); akan ada sekitar 200 kategori tetapi setiap pembeli bisa hanya dimuat di salah satu kategori. Satu batang tidak akan mencukupi, dan kadang-kadang mungkin ada perbedaan besar persentase antar negara, katakanlah, jumlah pelanggan dari Amerika Serikat dan yang lain dari uganda. Maka akan grafik batang Uganda sangat kecil sehingga tidak terlihat.



Gambar 4.1 Persentase jawaban terhadap pertanyaan biner.

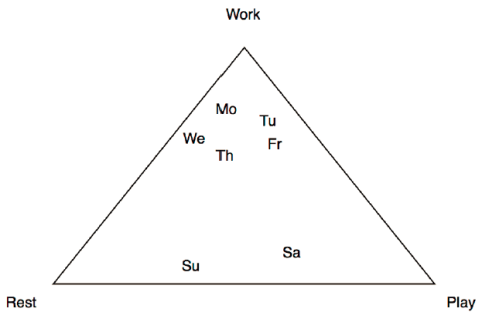
Sebelumnya telah disinggung tentang diagram lingkaran, tetapi grafik semacam wafel untuk dapat mengatasi masalah ini (Gambar 4.2). Tentu saja, kategori kecil dapat disatukan sebagai "Lainnya" dan

catatan kaki dapat menjelaskan apa yang dikandungnya. Jika terdapat sejumlah besar kategori, pembaca akan kesulitan menyerap semua informasi, dan mungkin yang terbaik adalah dengan menggabungkan kategori yang berisi sedikit data menjadi kategori khusus. Dengan pengujian akan dapat membantu menemukan ambang yang tepat untuk ini.



Gambar 4.2 Grafik Wafel 10-kali-10 menunjukkan persentase dari enam kategori dengan satu yang di*highlight* (Louisa Degenhardt dkk).

Ternary Plot adalah cara menunjukkan variabel yang berisi tiga kategori (Gambar 4.3). Jumalah masing-masing harus ditambah bersama hingga 100%. Plot seperti ini popular di dalam ilmu geologi, kimia dan makanan dan bidang-bidang ilmu yang berususan berurusan dengan campuran. Di setiap sudut segitiga, ditampilkan 100% salah satu kategori dan 0% untuk kategori lainnya. Pada Gambar 4.3, aktivitas di hari-hari dalam seminggu yang terdiri dari bekerja (*work*), bermain (*play*) dan istirahat (*rest*).

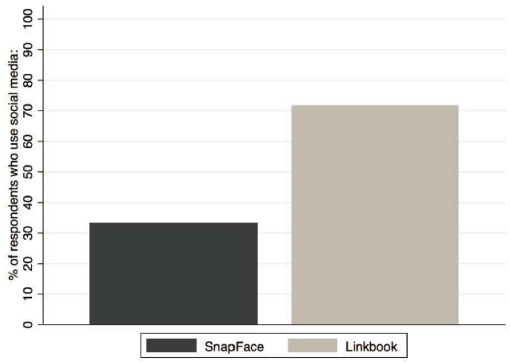


Gambar 4.3 *Ternary Plot* proporsi aktivitas setiap hari dalam satu minggu.

Ada dua kasus khusus untuk dipikirkan. Pertama, pertanyaan jenis “centang semua yang cocok", di mana persentase mungkin bertambah lebih dari 100%. Pada dasarnya, pertanyaan semacam ini harus dipecah menjadi serangkaian pertanyaan biner ya/tidak untuk masing-masing opsi. Kedua, variabel ordinal memiliki urutan bawaan, dan itu selalu harus dijaga dalam visualisasi. Hal ini menimbulkan pembatasan lain dalam visualisasi. Karena itu kompromi tidak bisa dihindari.

4.2 Perbandingan Data Tak Sepadan

Perbandingan data persentase dari dua variabel, dan disandikannya sebagai panjang dapat disajikan bersebelahan seperti dalam Gambar 4.4. Bentuk grafik area, atau *wafel plot* pun, tidak bisa dilihat secara visual untuk dibandingkan secara mudah oleh pembaca. Jika persentase ditambahkan dalam teks ke bagan seperti ini, maka akan mengacaukan gambar dan mengalihkan perhatian pembaca.

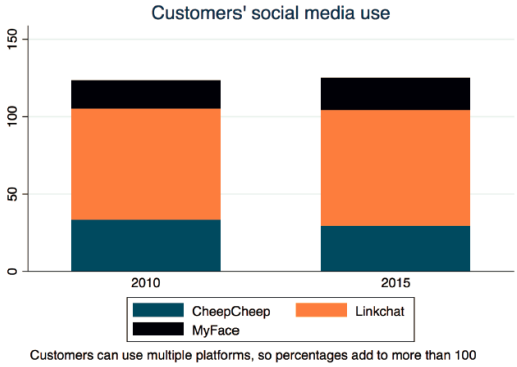


Gambar 4.4 Bagan perbandingan persentase dua jawaban

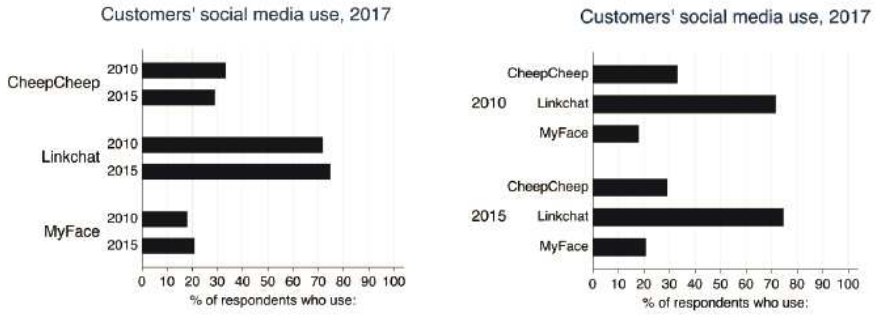
Jika terdapat banyak variabel atau kategori yang akan dibandingkan pada banyak titik waktu, maka bagan tetap dapat diatur meski membandingkannya semakin sulit. Karena itu ada yang harus dikorbankan, misalnya dengan mengelompokkan beberapa kategori tidak penting dalam satu obyek dan perbandingan dilakukan hanya pada kategori penting yang lain.

Opsi menumpuk grafik batang (Gambar 4.5) diberikan dalam banyak paket perangkat lunak analisis data dan spreadsheet, tetapi ini

bermasalah. Meskipun kategori di bagian bawah bilah bisa dibandingkan secara visual di seluruh bar, semua yang di bagian atasnya tidak bisa dibandingkan (tanpa menggunakan "pita pengukur") sperti yang ditunjukkan pada Gambar 2.7 dimana tidak masing-masing dimulai pada level yang berbeda.



Gambar 4.5 *Bar chart* bertumpuk, bandingkan dengan Gambar 4.6

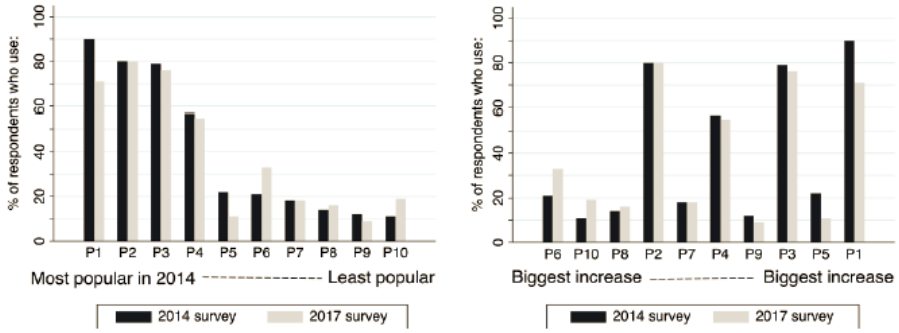


Gambar 4.6 *Bar chart* terklaster membandingkan tiga vaiabel binari menurut waktu, perhatikan perubahan kecil dapat dengan mudah terlihat di grafik sebelah kiri dibandingkan dengan grafik *Bar chart* bertumpuk di Gambar 4.5

4.3 Perbandingan Data Sepadan

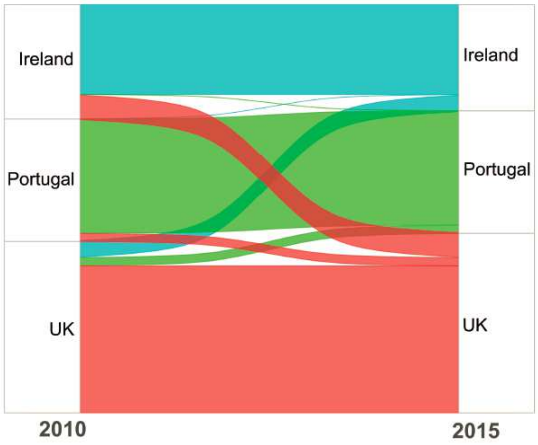
Dalam variabel kontinu dan diskrit di Bab 3, sering dapat ditemukan data sepadan. *Bar Chart* terklaster, *dot plot* dan line chart merupakan

kunci di sini (Gambar 4.6). Pembaca dapat terbantu melihat pertanyaan mana yang memiliki persentase tertinggi, atau perubahan terbesar, melalui pengaturan pasangan dalam urutan (Gambar 4.7). Ini adalah contoh sederhana untuk menekankan pesan tanpa mendistorsi fakta. Saat membandingkan titik waktu, bandingkan *like* dengan like. Sebagai contoh, jika responden untuk survei 2017 sangat berbeda dari survei 2014, pembaca perlu diberi tahu tentang hal itu di catatan kaki. Mungkin lebih baik dibandingkan hanya sebagian yang terjadi pada kedua kategori. Terkadang, pencocokan data dilakukan dengan cara lain selain waktu. Misalnya seorang dokter studi arthritis, dapat membandingkan pinggul yang terkena dengan pinggul sehat pada setiap responden. Ini masih sepadan dan masih cocok untuk ditampilkan seperti itu.



Gambar 4.7 Perbandingan sepubuh variabel binari, diurut menurut nilai (kiri) dan perubahannya (kanan).

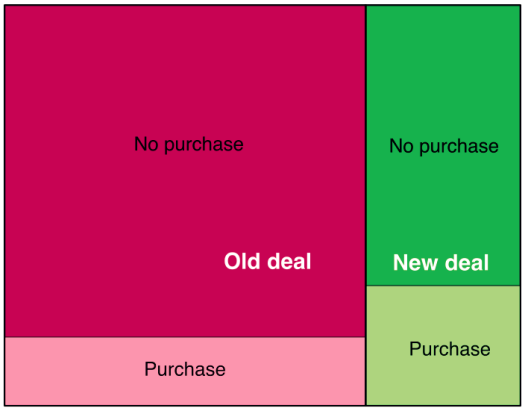
Terkadang, yang paling penting bukanlah angka pada suatu kategori tetapi data perubahannnya yang lebih penting. Ini membawa kita ke konsep *rasio* dan *perbedaan*. Kajian ini akan dibahas pada bab selanjutnya. Saat ini perhatikan Gambar 4.7. Apakah semua orang menanggapi pertanyaan P2 tetap sama? Grafik batang seperti ini tidak bisa digunakan untuk membantu menjawabnya. Salah satu cara menunjukkan aliran data yang cocok dengan hal ini adalah dengan *paralel set plot* (Gambar 4.8), meskipun tampak agak berantakan. Cara ini analogi dengan bagan garis pada Gambar 3.8 sebelumnya.



Gambar 4.8 Menampilkan pergerakan antar kategori pada data sepadan melalui *paralel set charts* - 2 orang telah pindah dari Irlandia ke Inggris, 1 dari Irlandia ke Portugal, 2 dari Inggris ke Irlandia, dan 1 dari Portugal ke Irlandia.

4.4 Kategori dalam Kategori

Ketika kategori bersarang di dalam kategori lain, maka akan didapatkan format pohon. Pohon keputusan dapat linier, dengan cabang-cabangnya selalu menuju ke bawah (atau ke atas, meskipun ke bawah tampaknya lebih umum) (Gambar 4.10). Bentuk bisa juga radial. Bentuk populer seperti ini misalnya adalah *donut chart*, diaman digunakan proporsi sudut (atau area); ring selanjutnya dapat berisi subdivisinya, dan ini disebut *sunburst chart*. *Treemap* merupakan pilihan lain (Gambar 4.9). Semua grafik golongan ini mengalami masalah jika pembaca harus membandingkan daerah/item yang tidak bersebelahan, dan item-item ini mungkin tidak memiliki titik awal yang sama (karena lokasi cabang yang berbeda). *Treemap* dan *donut graph* juga bergantung pada persepsi pembaca terhadap area.



Gambar 4.9 *Treemap*: percobaan pemasaran membandingkan dua penawaran dan apakah pengunjung situs web melakukan pembelian. Meskipun lebih sedikit pengunjung yang ditawari *new deal*, namun pemisahan yang terjadi lebih tinggi pada segmen hijau menjadi *purchace* dan *no purchace* dan ini menunjukkan *new deal* lebih populer dari pada *old deal*.

4.5 Asosiasi

Penggunaan asosiasi antara dua variabel sangat penting dalam statistik dan sain data. Sejauh ini, sudah dibahas beberapa cara membandingkan persentase pada kategori-kategori dan membandingkan kategori-kategori pada beberapa titik waktu. Berarti kombinasi dari satu variabel (kategori) dengan yang lain (waktu) dapat dilakukan secara umum untuk mengetahui apakah nilai satu variabel terkait dengan variabel yang lain.

Misalnya, bayangkan data yang diberikan oleh orang yang menerima narkoba perawatan untuk radang sendi. Mereka ditanya seberapa kepuasan mereka dengan obat yang didapat dari perawatan (“tidak puas ”atau“ puas ”) dan juga apakah mereka telah mencoba terapi pelengkap, seperti akupunktur. Mungkin persentase orang yang mencoba akupunktur lebih tinggi pada orang yang *tidak puas* daripada pada orang yang *puas*. Ini menunjukkan bukti bahwa obat yang diberikan tidak memadai dan menyebabkan orang mencari sendiri perawatan tambahan dalam bentuk akupunktur.

Kemungkan sebaliknya: pengalaman dirawat akupunktur dengan tidak tergesa-gesa memberi kesan ke pasien menjadi kurang puas terhadap konsultasi medis singkat. Atau terdapat penyebab umum: beberapa orang tidak percaya dengan obat konvensional. Atau mungkin tidak ada koneksi sebab-akibat yang bisa dimengerti, tetapi meskipun demikian data kurangnya kepuasan dapat diggunakan untuk mengidentifikasi orang yang mungkin ingin mencoba akupunktur.

Dalam kasus seperti ini, perlu dilihat persentase kondisional: persentase yang mencoba akupunktur pada kelompok "tidak puas" (atau disebut sebagai "tergantung pada kepuasan"). Untuk persentase bersyarat, ada beberapa opsi selain *clustered bar*.

Treemaps membagi area persegi panjang berdasarkan satu variabel dikodekan ke panjang horisontal dan yang lain kearah vertikal (Gambar 4.9); meskipun nampak kesan area tidak akurat, namun kesan akurat ada pada panjang di kedua sisi.

Pohon keputusan menunjukkan penguraian data oleh satu variabel kemudian oleh variabel yang lain dengan cara yang sangat intuitif, meskipun secara umumnya hanya diagram yang tidak benar-benar disandikan data secara visual. Namun, piktogram atau wafel dapat ditumpangkan di tiap titik di mana pohon keputusan bercabang atau berakhir, sekilas memberi kesan untuk angka-angka dan bagaimana mereka menyebar melalui pohon (Gambar 4.10). Jumlahnya bisa berupa pengamatan nyata atau hipotesis: misal departemen pemasaran perusahaan teknologi mungkin menunjukkan berbagai skenario tempat pengguna berpindah dari mengunjungi situs web mereka ke mengunduh aplikasi untuk melakukan pembelian dalam aplikasi.

Gambar 4.10 *Decision Tree* digabung dengan grafik *wafel*, membandingkan dua *deal* apakah pengunjung melakukan pembelian,

Idealnya, pembaca harus mampu mengolah persentase secara mental sesuai minat mereka. Telah ditunjukkan cara bagaimana mevisualisai data pengunjung situs web yang ditawarkan melakukan pembelian dan cara mengetahui berapa banyak dari mereka yang melakukan pembelian. Visualisasi harus memudahkan pembaca bahkan tanpa menghitung kotak *wafel* di atas.

Penelitian kualitatif digunakan dalam banyak latar belakang, baik ilmiah atau komersial, untuk mencri pola dalam data. Relevansi hal ini dalam visualisasi data adalah bahwa vaisualisai harus dapat menyampaikan pesan sampai batas tertentu. Dalam kasus asosiasi antar variabel, pesan itu terkadang menjadi hilang. Jadi dalam visualisasi, berhati-hatilah untuk tidak memaksakan interpretasi yang terlalu banyak ketika menunjukkan asosiasi. Jika membaca visualisasi, jangan menganggap pesan yang menyertai data sebagai data. Bab ini adalah telah membahas data, statistik, dan persentase. Selanjutnya akan dibahas visualisasi model prediksi pada bab-bab selanjutnya.

4.5 Risiko dan Laju, dan Peluang

Data risiko sebenarnya merupakan data proporsi atau persentase seperti yang data lain, tetapi penyebutannya harus hati-hati, baik dalam gambar atau dalam teks yang menyertainya. Apa yang terjadi ketika risikonya tidak lagi menjadi perbandingan yang adil? Jika departemen pemasaran hanya mempertimbangkan telah melacak pelanggan untuk waktu yang lama dan menghitung apakah seseorang telah membuat pembelian atau tidak, maka itu mungkin bukan perbandingan yang adil untuk melihat pengunjung lama dengan cara yang sama seperti melihat pendatang baru. Pasti pendatang baru cenderung melakukan pembelian. Dalam kasus seperti ini, data jumlah pembelian harus mempertimbangkan lamanya seseorang mengunjungi situs web. Ini memberi kita *rate* (laju/tingkatan), dan meskipun harga tidak masuk dalam kisaran 0% hingga 100%, teknik yang sama yang telah kita lihat di bab ini bisa memvisualisasikannya.

Terkadang, untuk alasan komputasi, kita harus membicarakannya peluang terjadinya suatu peristiwa dan bukan risiko. Kemungkinannya adalah hanya jumlah jawaban "ya" (atau setara) yang dibagi dengan jumlah jawaban "tidak" (bukan total). Dalam contoh pemasaran di atas, kemungkinan melakukan pembelian jika Anda ditawari *new deal* adalah 270/630 = 0,43. Peluang tidak pernah ditulis sebagai persentase (43%) untuk menghindari kebingungan. Hindari menggunakan peluang dalam visualisasi data, karena terbukti membingungkan pembaca. Gunakan data persentase untuk lebih jelasnya.