

6. Analisis Statistik

6.0 Kebarangkalian

6.1 Taburan statistik

6.2 Taabiran statistik

6.3 Anggaran biparameter

6.4 Anggaran berparameter

- Tugas 12

6.0 Kebarangkalian

6.0.0 Asas

Suatu set peristiwa $E \in \Omega$ berada dalam ruang kebarangkalian Ω . Maka kebarangkalian $P(E)$ ialah kekerapan mendapat hasil E dalam cubaan yang banyak. Ini ialah definisi kebarangkalian berdasarkan kekerapan. Kebarangkalian $P(A) = P(x \in A)$ di mana x adalah pembolehubah rawak. Pembolehubah rawak boleh samada bersifat diskret atau selanjar.

Aksiom kebarangkalian meyakinkan jumlah kebarangkalian

$$P(\Omega) = 1,$$

dan nilai kebarangkalian,

$$0 \leq P(E) \leq 1.$$

Bagi set peristiwa-peristiwa A_i yang tak bercantum, kebarangkalian jumlah

$$P(\cup A_i) = \sum P(A_i).$$

Dengan itu, kita perolehi teorem-teorem berikut,

$$P(A) = 1 - P(A^*)$$

di mana A^* ialah pelengkap kepada A , $\Omega = A \cup A^*$. Juga,

$$P(\emptyset) = 0$$

Bagi $A_1, A_2 \subseteq \Omega$ dan $A_1 \subset A_2$, maka

$$P(A_1) \leq P(A_2).$$

Bagi $A_1, A_2 \subseteq \Omega$ yang tak semestinya tercantum,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

6.0.1 Fungsi taburan

Suatu fungsi taburan kebarangkalian ("pdf") ditulis $f(\omega)$.

Fungsi taburan melonggok atau fungsi taburan kumulatif adalah

$$F(x) = \sum_{\omega \leq x} f(\omega)$$

bagi ω diskret, atau

$$F(x) = \int_{-\infty}^x f(\omega) d\omega$$

bagi ω selanjar.

Bagi suatu pembolehubah rawak x dengan taburan $u(x)$, nilai jangkaan ialah fungsian terhadap fungsi itu,

$$E[u(x)] = \int_{-\infty}^{\infty} u(x) f(x) dx$$

bagi x selanjar. Fungsian ini mempunyai ciri-ciri berikut,

$$E[k] = k$$

jika k pemalar. Kemudian,

$$E[kv] = k E[v]$$

dan

$$E[k_1 v_1 + k_2 v_2] = k_1 E[v_1] + k_2 E[v_2]$$

yang menyatakan E itu linear. Maka,

$$E[x] = \int x f(x) dx = \int x dF(x).$$

Perlu dihargai bahwa nilai jangkaan kadangkala tak wujud.

Median ialah nilai pertengahan dalam suatu set data yang diatur. Ini bermakna, jika x_M itu nilai median,

$$\int_0^{x_M} f(x) dx = \int_{x_M}^{\infty} f(x) dx,$$

ataupun,

$$F(x_M) = 1/2.$$

Mod pula ialah nilai yang paling berkemungkinan, iaitu x bila $f(x)$ maksimum.

Pengukuran cirian suatu taburan boleh dilakukan oleh nilai momennya. Momen ke- m bagi kuantiti x untuk $f(x)$ ialah

$$E(x^m) = \int_{\Omega} x^m f(x) dx$$

untuk x selang atau

$$E(x^m) = \sum_x x^m f(x)$$

untuk x diskret. Perhatikan bahawa min adalah momen pertama

$$\mu = E(x),$$

yang merupakan ukuran kedudukan.

Momen diukur terhadap suatu titik. Momen pusatan ialah momen terhadap min,

$$E[(x - \mu)^m] = \int_{-\infty}^{\infty} (x - \mu)^m f(x) dx$$

Lihat

$$E[x - \mu] = 0.$$

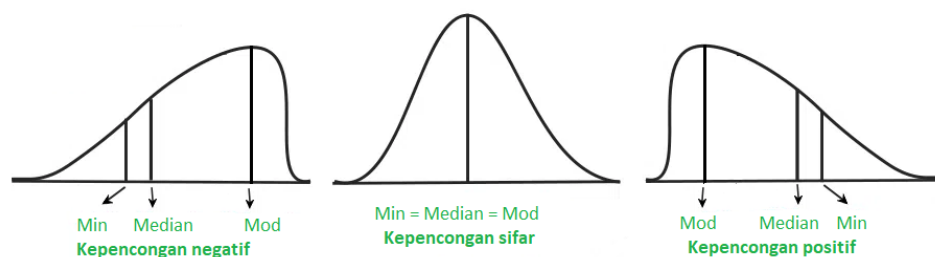
Varians,

$$\sigma^2 \equiv V(x) = E[(x - \mu)^2] = \dots = E[x^2] - E[x]^2$$

mengukur kelebaran taburan. Kepencongan pula diberikan oleh,

$$\gamma_1 = E[(x - \mu)^3] / \sigma^3.$$

Ia ukuran tak seimbangny sesuatu taburan itu.



Kurtosis,

$$\gamma_2 = \{E[(x - \mu)^4] - 3\} / \sigma^4,$$

mengukur bagaimana sesuatu taburan itu 'tajam puncak' (kurtosis positif) atau 'berekor' (kurtosis negatif) berbanding taburan Gaussian atau normal (kurtosis sifar).

Kita boleh ada keadaan taburan kebarangkalian mempunyai lebih daripada satu pembolehubah.

Dalam kes dua pembolehubah x_1 dan x_2 misalnya, kita tuliskan taburan itu sebagai $f(x_1, x_2)$.

Kalau x_1 dan x_2 bebas, maka dan jika boleh ditulis $f(x_1, x_2) = g(x_1)h(x_2)$. Korelasi di antara pembolehubah boleh diukur oleh kovarians,

$$\text{cov}(x,y) = E[(x-\mu_x)(y-\mu_y)] = E(xy) - \mu_x\mu_y$$

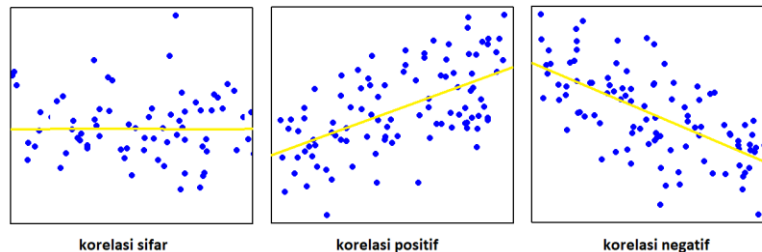
Perhatikan bahawa, jika x, y bebas, $E(xy) = E(x)E(y)$ maka kovarians adalah 0. Pekali korelasi, ialah

$$\rho_{x,y} = \frac{\text{cov}(x,y)}{\sigma_x\sigma_y}$$

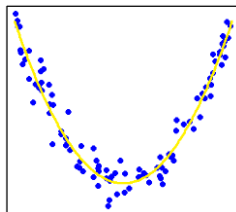
di mana sisihan piawai ialah punca kuasadua varians,

$$\sigma_x = +\sqrt{E[(x-\mu_x)^2]},$$

dan lain-lainnya. Rajah di bawah menunjukkan taburan dua pembolehubah x, y , dengan pelbagai korelasi (garisan untuk panduan mata):



Namun, nilai korelasi 0 tidak semestinya bermakna tiada pergantungan di antara pembolehubah. Suatu contoh ditunjukkan di bawah:



6.0.2 Hukum

Hukum lemah Nombor Besar mengatakan,

dalam had $n \rightarrow \infty$,

$$\frac{1}{n} \sum_i x_i \rightarrow \mu.$$

Dalam had bilangan besar data, nilai purata menumpu kepada nilai min.

Teorem Had Pusat pula mengatakan,

$$\lim_{n \rightarrow \infty} P(z_n \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{1}{2}u^2} du \equiv \text{erf}z$$

(Ada juga definisi fungsi ralat

$$\text{erf}z \equiv \frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} du$$

yang dipakai.) Taburan kumulatif taburan normal ialah

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

Teorem ini menyatakan, purata (atau jumlah) pembolehubah rawak

$$y_n = \frac{1}{n} \sum_i x_i$$

yang bebas dan teragih serupa akan mendekati taburan normal atau taburan Gauss apabila bilangan pembolehubah itu menjadi besar, walaupun ia asalnya tidak tertabur secara normal. Taburan purata sampel akan mendekati taburan normal walaupun taburan asal tidak normal. Ini memberikan asas yang kuat untuk menggunakan kaedah statistik berasaskan taburan normal (seperti ujian-t, selang keyakinan, dan sebagainya), kerana walaupun data asal tidak teragih secara normal, teorem ini mengatakan bahawa purata sampel besar akan teragih hampir normal.

6.0.3 Kebarangkalian bersyarat

Tulis kebarangkalian bagi E jika diberi A benar sebagai $P(E|A)$.

Maka

$$P(A|A) = 1$$

dan

$$P(A_2|A_1) = P(A_1 \cap A_2 | A_1).$$

Oleh kerana

$$\frac{P(A_1 \cap A_2 | A_1)}{P(A_1 | A_1)} = \frac{P(A_1 \cap A_2)}{P(A_1)}$$

maka

$$P(A_2 | A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)}$$

Fungsi taburan taburan kebarangkalian bersyarat, ialah,

$$f(x_2 | x_1) = \frac{f(x_1, x_2)}{f(x_1)}$$

Teorem Bayes menyatakan, bagi B_i saling eksklusif, $\sum_i B_i = 1$,

$$P(A) = \sum_i P(A|B_i) P(B_i),$$

jadi

$$P(B_i | A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

posteriori.

6.0.4 Penukaran pembolehubah

Katakan pembolehubah rawak x tertabur dengan pdf $f(x)$, dan $y=u(x)$ merupakan transformasi 1-ke-1 x ke atas y . Maka songsangan,

$$x = \omega(y).$$

Fungsi taburan kebarangkalian untuk y :

$$g(y) = P(y) = P(x=\omega(y)) = f[\omega(y)].$$

Bagi kes multivariat, bagi transformasi $(x_1, x_2) \rightarrow (y_1, y_2)$,

$$y_1 = u_1(x_1, x_2)$$

$$y_2 = u_2(x_1, x_2)$$

songsangan adalah,

$$x_1 = \omega_1(y_1, y_2)$$

$$x_2 = \omega(y_1, y_2)$$

Boleh ditulis

$$f(x_1, x_2) dx_1 dx_2 = f(\omega_1(y_1, y_2), \omega(y_1, y_2)) |J| dy_1 dy_2$$

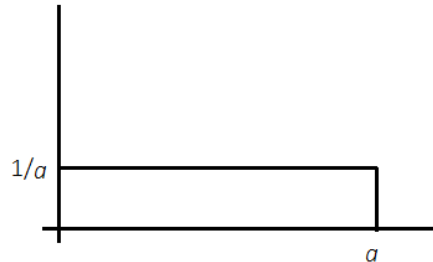
di mana

$$J = \begin{vmatrix} d\omega_1/dy_1 & d\omega_1/dy_2 \\ d\omega_2/dy_1 & d\omega_2/dy_2 \end{vmatrix}$$

adalah Jacobian untuk transformasi berkenaan.

6.1 Taburan statistik

6.1.0 Taburan Seragam



Bagi taburan seragam,

$$U(x) = 1/a \quad 0 \leq x \leq a$$

$$E(x) = a/2$$

$$V(x) = a^2/12$$

6.1.1 Taburan Bernoulli

Ini diberikan oleh proses Bernoulli yang merupakan jujukan di antara 2 kesudahan:

$$k = 0 \quad P(k) = q \quad \text{'gagal'}$$

$$k = 1 \quad P(k) = p = 1 - q \quad \text{'berjaya'}$$

Maka

$$E(x^m) = E(x) = p$$

$$V(x) = p - p^2 = pq$$

Bagi n cubaan bebas, untuk k kejayaan,

$$P(k, n) = \binom{n}{k} p^k q^{n-k}$$

di mana pekali binomial

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Inilah taburan Bernoulli.

Taburan binomial $B(k, n, p) = P(k, n)$ bila kebarangkalian kejayaan individu = p . Maka

$$E(k) = nE(B(1, 1, p)) = np$$

$$\sigma_k^2 = npq = np(1-p)$$

Kemudian, taburan multinomial

$$M(x_i, p_i, n) = \frac{n!}{x_1! x_2! \dots x_n!} p_1^{x_1} \dots p_k^{x_k}$$

Satu contoh taburan multinomial ialah histogram:

n = bil. Peristiwa

x_i = isi bin ke- i ; $\sum p_i = 1$

Untuk ini,

$$E(x_i) = np_i$$

$$V(x_i) = \sigma_i^2 = np_i(1-p_i)$$

Sila perhatikan bahawa x_i tak semua bebas kerana jumlahnya $= n$.

6.1.2 Taburan Poisson

Taburan ini bercirikan:

- bilangan peristiwa dalam selang berlainan, adalah bebas
- bilangan peristiwa berkadar saiz selang dalam had saiz ini $\rightarrow 0$
- untuk kebarangkalian yang > 1 peristiwa/selang, ia $\rightarrow 0$ bila saiz selang $\rightarrow 0$
- ini semua bebas kedudukan, iaitu

$$P_{\alpha t}(n) = \frac{e^{-\alpha t} (\alpha t)^n}{n!}$$

di mana α = kadar. Jadi

$$E(n) \equiv \mu = \dots = \alpha t$$

dan

$$V(n) = \dots = \mu$$

- hasil tambah dua pembolehubah Poisson juga Poisson

6.1.3 Taburan Gaussian atau Normal

Taburan ini diberikan,

$$N(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

atau secara lebih am,

$$N(x; \mu, \sigma) = \frac{1}{\sigma} N\left(\frac{x-\mu}{\sigma}; 0, 1\right)$$

Adalah

$$\int_{-\infty}^z N(x; \mu, \sigma) dx = \text{erf}\left(\frac{z-\mu}{\sigma}\right)$$

Momen pusat

$$E[(x-\mu)^n] = 0$$

bagi n ganjil, dan

$$E[(x-\mu)^n] = (2k-1)!! \sigma^{2k}$$

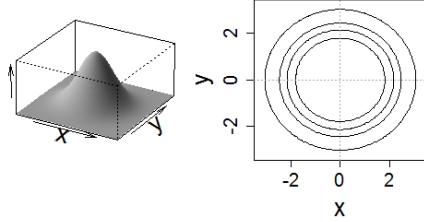
untuk $n = 2k$.

Dwivariat Normal diberikan:

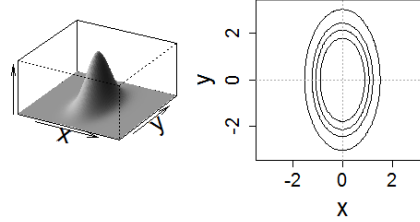
$$f(x, y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2(1-\rho^2)} \left[\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right]\right\}.$$

bagi pembolehubah x, y . Gambarajah berikut menunjukkan beberapa kes untuk berlainan σ_x, σ_y .

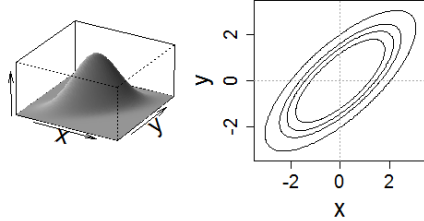
$$\sigma_x = \sigma_y, \rho = 0$$



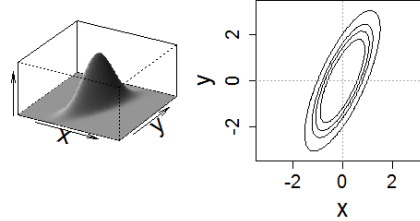
$$2\sigma_x = \sigma_y, \rho = 0$$



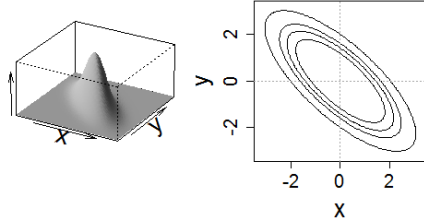
$$\sigma_x = \sigma_y, \rho = 0.75$$



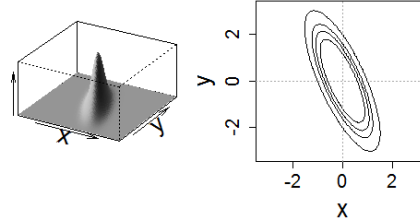
$$2\sigma_x = \sigma_y, \rho = 0.75$$



$$\sigma_x = \sigma_y, \rho = -0.75$$



$$2\sigma_x = \sigma_y, \rho = -0.75$$



Matriks Kovarians bagi taburan dwivariat diberikan,

$$V = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

Jika V berpepenjuru, ia bermakna x, y adalah bebas. Boleh didefinisikan

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

Bagi taburan dwivariat, boleh cari pertukaran pebolehubah (yang merupakan kombinasi linear pembolehubah yang lama) sehinggakan ia bebas atau tak berkorelasi.

Multivariat Normal (kordinat $x_1, x_2, \dots, x_n$) punyai matriks kovarian

$$V =$$

$$\begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots \\ \vdots & \ddots & \\ \rho_{n1}\sigma_n\sigma_1 & \dots & \sigma_n^2 \end{bmatrix}$$

6.1.4 Taburan χ^2

Jika $x_1, x_2, \dots, x_n$ merupakan pembolehubah rawak normal bebas (n = darjah kebebasan) maka

$$f(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = N(x_1; \mu_1, \sigma_1) N(x_2; \mu_2, \sigma_2) \dots N(x_n; \mu_n, \sigma_n)$$

dan didefinisikan

$$\chi^2 = \sum_{i=1}^n \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2$$

Taburan χ^2 ini digelar taburan χ^2 . Minnya,

$$\langle \chi^2(n) \rangle = n,$$

sementara variansnya,

$$\sigma_{\chi^2(n)}^2 = 2n.$$

6.1.5 Taburan Cauchi

Taburan ini diberikan,

$$C(x; \mu, \alpha) = \frac{1}{\pi\alpha} \frac{1}{1 + \frac{(x-\mu)^2}{\alpha^2}}$$

Variansnya ∞ dan ia tak taat hukum nombor besar.

6.2 Taabiran Statistik

6.2.0 Asas

Taabiran statistik ialah proses seperti berikut:

Populasi → (penyampelan) → data → (statistik) → taabiran tentang populasi
Kesimpulan statistik tentang populasi dibuat daripada data yang disampel daripada populasi.

Jenis utama kesimpulan statistik adalah:

- Anggaran
Menganggar nilai atau julat nilai sesuatu ciri populasi berdasarkan yang berkenaan bagi suatu sampel
- Ujian hipotesis
Menilai sesuatu idea berkenaan sesuatu populasi menggunakan suatu sampel

6.2.1 Anggaran

Nilai yang paling mungkin bagi sesuatu parameter populasi dipanggil anggaran titik. Dalam menganggarkan nilainya, selalu ada ketidakpastian..Ketidakpastian sering dinyatakan dalam bentuk selang keyakinan yang ditakrifkan oleh nilai terendah dan tertinggi yang mungkin untuk parameter berkenaan. Contoh, "Purata bilangan basikal yang dimiliki oleh orang Besut adalah antara 3.5 dan 6."

6.2.2 Ujian hipotesis

Sesuatu hipotesis dibuat tentang populasi. Contoh, "Purata bilangan basikal yang dimiliki oleh orang Besut adalah 4.8." Ujian hipotesis ialah untuk menilaikan kebarangkalian bahawa hipotesis adalah benar, berdasarkan data sampel.

Ada pelbagai jenis ujian hipotesis. Langkah-langkah ujian bergantung kepada jenis data (sama ada ia berbentuk kategori atau berangka), dan domainnya (sama ada satu kumpulan, atau bandingan dua kumpulan, atau kumpulan sama, sebelum-selepas). Lagi contoh hipotesis atau soalan yang diuji:

70% rakyat Malaysia kidal.

Adakah berat purata kucing kurang daripada 5 kg?

Adakah pendapatan doktor lebih daripada pendapatan peguam?

6.2.3 Statistik

Disebut 'statistik' ialah apa-apa fungsi cerapan yang tidak bergantung kepada parameter populasi yang tidak diketahui.

6.2.4 Penganggar

Penganggar adalah apa-apa statistik yang digunakan untuk meneka nilai suatu kuantiti, katakan θ . $\hat{\theta}$ adalah anggaran titik untuk θ .

Ciri-ciri baik untuk penganggar adalah seperti berikut:

- Tak pincang – sentiasa berikan nilai yang betul dalam eksperimen yang banyak
- Varians minimum
- Konsisten – dengan bertambah sampel, penganggar menumpu ke nilai yang betul
- Mencukupi – menyimpan semua maklumat berkenaan θ yang ada dalam sampel
- Teguh – tak sensitif kepada perubahan dalam model andaian
- Kehilangan maklumat minimum
- Kos pengiraan minimum

Penganggar bagi kedudukan dalam taburan, misalnya jisim zarah dalam taburan ukuran jisim bersama ralat, mashayat zarah dalam taburan reputan. antaranya, adalah

$$\hat{\mu} = \bar{x} \quad \text{min}$$

$$\hat{\mu} = \text{median} \quad (\text{berguna dalam keadaan taburan dengan ekor panjang})$$

$$\hat{\mu} = \frac{1}{2} [\min(x_i) + \max(x_i)] \quad (\text{berguna untuk taburan seragam tanpa ekor})$$

$$\hat{\mu} = \bar{x} \text{ terpapas} - \text{ekor dicantas}$$

Penganggar tak terpincang ada punyai masalahnya:

- Tak sentiasa mudah
- Mungkin tak wujud
- Mungkin ralat min kuasaduanya bukan yang terkecil
- Ketakpincangan tak semestinya takvarian terhadap pertukaran pembolehubah

Berikut adalah beberapa cara membina penganggar.

Dalam kaedah penggantian frekuensi, digantikan frekuensi populasi yang tak diketahui dengan frekuensi sampel. Sebagai contoh, dikehendaki $g(p)$ untuk taburan binomial $B(x, n, p) = \binom{n}{p} p^x (1-p)^{n-x}$. Gantikan p dengan x/n kerana $E(x) = np$. Kelebihannya ialah ia mudah sambil konsisten. Kekurangannya ialah ia mungkin pincang, dan mungkin variansnya bukan terkecil.

Begini ialah cara pengurangan pincang bagi penganggar. Katakan $\hat{\theta}$ adalah penganggar takpincang untuk θ . Kita mahu anggar $g(\theta)$. Cuba $\hat{g} = g(\hat{\theta})$, yg mungkin ada sedikit pincang. Kembangkan sebagai siri Taylor,

$$g(\theta) = g(\hat{\theta}) + (\hat{\theta} - \theta)g'(\hat{\theta}) + \frac{1}{2}(\hat{\theta} - \theta)^2 g''(\hat{\theta}) + \dots$$

Maka

$$E[g(\theta)] \cong g(\hat{\theta}) + \frac{1}{2}V(\hat{\theta})g''(\hat{\theta}).$$

Bagi V kecil, maka,

$$\hat{g}_1(\theta) = \hat{g} - \frac{1}{2}v(\hat{\theta})g''(\hat{\theta}).$$

Untuk menganggarkan suatu fungsi kepada $q(\theta)$ menggunakan kaedah momen, ungkapkan $q(\theta)$

sebagai fungsi kepada momen-momen populasi,

$$m_j \equiv E(x^j),$$

dan kemudian gantikan momen populasi dengan momen-momen sampel. Yakni,

$$q(\theta) = g(m_1, m_2, \dots, m_n)$$

dengan $m_1 = \text{min}$, $m_2 = E(x^2) = v(x) + m_1^2$, dan seterusnya. Contoh, anggarkan

$$\sigma^2 \equiv m_2 - m_1^2.$$

Cuba

$$\hat{\sigma}^2 = \hat{m}_2 - \hat{m}_1^2 = \frac{1}{n} \sum x_i^2 - \bar{x}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

Kaedah momen ini tidak begitu teguh kerana momen-momen tinggi banyak bergantung kepada ekor-ekor dalam taburan.

6.3 Anggaran Biparameter

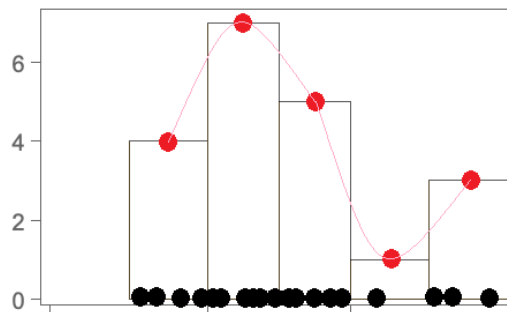
6.3.0 Asas

Kaedah-kaedah biparameter atau tanpa parameter membuat andaian minimum terhadap taburan populasi. Oleh kerana tiada taburan tertentu diandaikan bagi kebarangkalian data-data, kaedah-kaedah ini boleh lentur dan tegap dalam menangani data nyata. Ia tidak bergantung kepada kenormalan taburan atau kehomogenan varians. Kaedah biparameter adalah dipacu data, ia terfokus kepada data dan tidak bergantung kepada parameter-parameternya seumpama min dan varians. Data kecil, data tak seimbang, dan data dengan unsur luar, boleh ditangani.

Anggaran biparameter mencuba menganggarkan taburan kebarangkalian dasar bagi suatu set nilai-nilai rawak. Kita lihat tiga kaedah lazim, penghistograman, kaedah jiran terdekat, dan kebolehjadian maksimum.

6.3.1 Penghistograman

Domain set nombor-nombor rawak dibahagikan kepada selang-selang (disebut bin-bin). Biasanya saiz selang ini sama, namun ia juga boleh berbeza-beza. Bilangan nombor rawak yang jatuh ke dalam setiap selang dihitung, dan bilangan ini, dibahagi saiz selang, apabila ternormal, menganggarkan ketumpatan kebarangkalian taburan dasar, pada kedudukan di pertengahan selang. Suatu contoh mudah dilakarkan di bawah, dengan titik-titik hitam sebagai data nombor-nombor rawak, dan titik-titik merah sebagai titik-titik dalam keluk anggaran. Taburan secara selanjar dianggarkan oleh bentuk histogram.



Jika dalam bin ke- i , yang bersaiz s_i dan berpusat di x_i , ada n_i bilangan masukan data, taburan kebarangkalian dianggarkan

$$f(x_i) \propto n_i/s_i.$$

Pernormalan menghendaki $\int f(x_i) dx = \sum f(x_i) s_i = 1$. Jadi,

Penganggaran lebih baik diperolehi apabila sampel lebih besar, atau data lebih banyak.

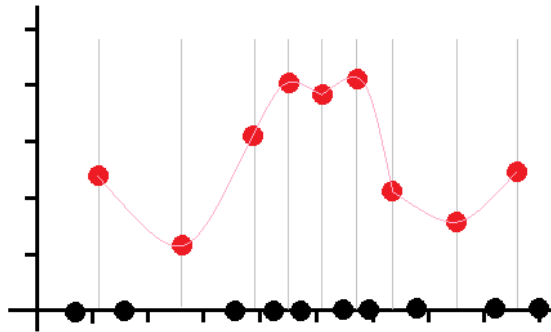
Walau apapun taburan kebarangkalian keseluruhan yang mendasar data, bilangan titik data dalam sesuatu bin itu punyai kerawakan dijangka seperti hingar tembakan (lihat bahagian 17.0) yang diperihalkan taburan Poisson. Min ukuran di bin i ialah n_i , maka oleh itu variansnya juga n_i . Jadi, ralat statistik bagi ukuran n_i ialah $\sqrt{n_i}$. Nisbah ralat lebih kecil bila lebih banyak titik data.

6.3.2 Kaedah jiran terdekat

Kalau penghistogramman menganggar ketumpatan menerusi hitungan titik dalam suatu selang, kaedah jiran terdekat membuatnya daripada jarak di antara titik data. Songsangan jarak ini memberikan ketumpatan. Ini jelas bagi taburan dalam 1 dimensi. Konsep ini boleh diitlakkan ke dimensi lebih tinggi dengan mudah.

Kaedah ini memberikan taburan yang licin jika titik-titik data tertabur secara seragam. Bagi nombok-nombor rawak, ini tidak benar. Untuk melicinkan lekuk anggaran, boleh digunakan jarak ke jiran terdekat ke- k dan bukan yang pertama. Begini, naik-turun nilai jarak dipuratakan untuk beberapa bilangan jiran.

Di bawah ini ditunjukkan suatu contoh anggaran jiran terdekat 1 dimensi bagi sampel titik-titik data (htiam), dan songsangan jarak di antara jiran terdekat (merah), diplot di kedudukan tengah-tengah di antara dua jiran berkenaan.



Nilai y fungsi anggaran kemudian perlu dinormalkan untuk memberikan jumlah kebarangkalian sifar.

6.3.3 Anggaran taburan kernel

Kaedah ini ialah itlakan kepada kaedah penghistogramman, Anggaran fungsi kebarangkalian ditulis sebagai hasil tambah kernel,

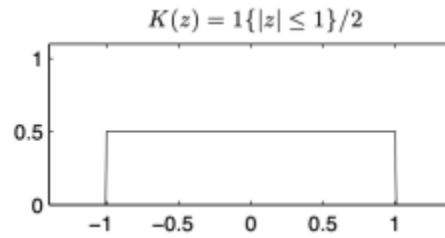
$$\hat{f}(x_0) = \frac{1}{Nh} \sum_i K\left(\frac{X_i - x_0}{h}\right)$$

di mana N ialah jumlah bilangan titik data dalam sampel. Fungsi kernel K mengukur dekatnya titik data rawak X_i kepada titik kedudukan x_0 , terhadap selang h , dan diberi pemberat sewajarnya. Ada beberapa jenis kernel dicadangkan.

Kernel seragam memberikan pemberat sama $\frac{1}{2}$ kepada setiap $X_i \in [x_0 - h, x_0 + h]$,

$$K\left(\frac{X_i - x_0}{h}\right) = \begin{cases} \frac{1}{2} & \text{if } \left|\frac{X_i - x_0}{h}\right| \leq 1 \\ 0 & \text{if } \left|\frac{X_i - x_0}{h}\right| > 1 \end{cases}$$

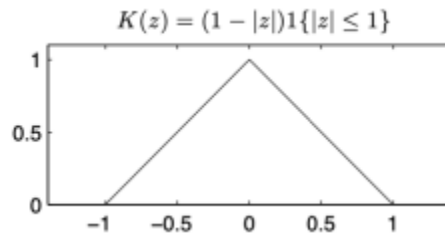
seperti di bawah,



Kernel tigasegi pula memberikan pemberat positif bagi $X_i \in [x_0 - h, x_0 + h]$, dengan nilai lebih tinggi jika X_i lebih dekat kepada x_0 ,

$$K\left(\frac{X_i - x_0}{h}\right) = \begin{cases} \left(1 - \left|\frac{X_i - x_0}{h}\right|\right) & \text{if } \left|\frac{X_i - x_0}{h}\right| \leq 1 \\ 0 & \text{if } \left|\frac{X_i - x_0}{h}\right| > 1 \end{cases}$$

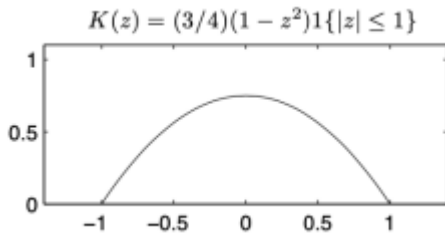
seperti di bawah,



Kernel Epanechnikov seperti kernel tigasegi, tetapi bentuk tigasegi diganti dengan bentuk melengkung,

$$K\left(\frac{X_i - x_0}{h}\right) = \begin{cases} \frac{3}{4} \left(1 - \left(\frac{X_i - x_0}{h}\right)^2\right) & \text{if } \left|\frac{X_i - x_0}{h}\right| \leq 1 \\ 0 & \text{if } \left|\frac{X_i - x_0}{h}\right| > 1 \end{cases}$$

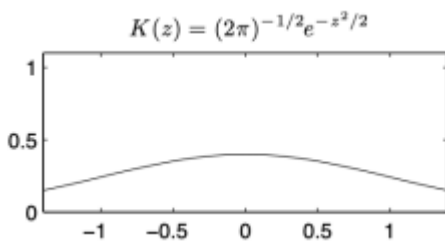
seperti di bawah,



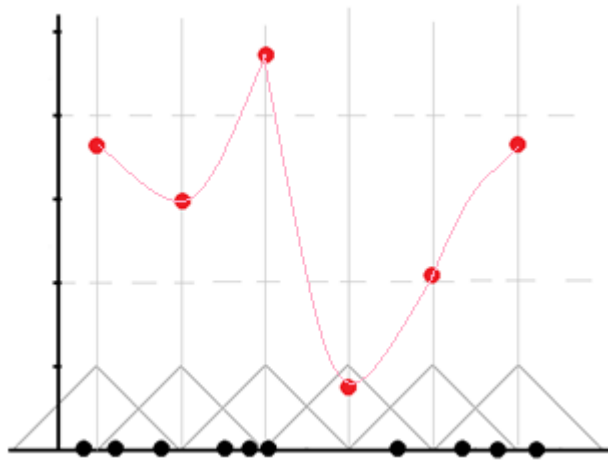
Kernel normal seterusnya adalah kernel yang menggunakan bentuk fungsi normal,

$$K\left(\frac{X_i - x_0}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left|\frac{X_i - x_0}{h}\right|^2}$$

seperti di bawah,



Contoh di bawah ini ialah titik di atas lengkung anggaran (merah) bagi kernel tigasegi ke atas titik-titik rawak sampel (hitam).



Perhatikan bahawa penghistograman dan jiran terdekat merupakan kaedah taburan kernel dengan kernel berkenaan masing-masing.

Boleh ditunjukkan, pincang

$$b(\hat{f}(x_0)) \equiv E(\hat{f}(x_0)) - f(x_0) = \frac{h^2}{2} f''(x_0) \int z^2 K(z) dz + O(h^4)$$

di mana $z = (x - x_0)/h$, dan varians

$$V(\hat{f}(x_0)) = \frac{1}{Nh} f(x_0) \int K(z)^2 dz + O\left(\frac{1}{Nh}\right).$$

Perhatikan tolak-ansur varians-pincang terhadap h – bila h kecil (model lebih bolehlentur dan kurang licin) mengurangkan pincang tetapi meningkatkan varians, dan sebaliknya bila h besar.

Ralat kuasadua min,

$$MSE(\hat{f}(x_0)) \equiv V(\hat{f}(x_0)) + b(\hat{f}(x_0))^2.$$

Kernel Epanechnikov meminimumkan MSE.

Semak kekonsistenan kaedah taburan kernel. Bila $N \rightarrow \infty$, $h \rightarrow 0$ dan $Nh \rightarrow \infty$,

$$b(\hat{f}(x_0)) \rightarrow 0$$

$$V(\hat{f}(x_0)) \rightarrow 0$$

$$\hat{f}(x_0) \rightarrow f(x_0).$$

6.4 Anggaran Berparameter

6.4.0 Asas

Kaedah anggaran berparameter ialah proses mengira nilai parameter sesuatu model dari data ukuran. Data sampel digunakan untuk menganggarkan parameter sesuatu taburan.

Ini bermakna perihalan analisan taburan telah dikenalpasti, dan hanya nilai parameter-parameternya harus ditentukan supaya lengkung perihalan yang ditala menyetujui sampel data yang diberi/diukur.

Data x diberikan taburan $f(x, \theta)$ dengan set parameter $\theta = (\theta_1, \dots, \theta_k)$.
 $\mathbf{x} = (x_1, \dots, x_n)$ adalah sampel.

6.4.1 Kebolehjadian maksimum

Fungsi taburan kebarangkalian gabungan ialah $g(\mathbf{x}, \theta)$. Jika x_i bebas, kebarangkalian boleh dihasildarabkan, jadi fungsi taburan kebarangkalian gabungan ialah $\prod_i f(x_i, \theta)$. Diberi sampel data $\mathbf{x}$, fungsi kebolehjadian terhadap θ ialah

$$\mathcal{L}(\theta) = \prod_i f(x_i, \theta).$$

Prinsip kebolehjadian maksimum ialah bahawa anggaran terbaik bagi satu set data ialah yang diberikan oleh nilai-nilai θ yang memaksimumkan fungsi kebolehjadian. Maksimum $\mathcal{L}$ sepadan dengan maksimum $\ln \mathcal{L} \equiv \mathcal{L}$. Ini berlaku bila

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{1}{\mathcal{L}} \frac{\partial \mathcal{L}}{\partial \theta} = 0$$

Kita gunakan $\mathcal{L}$ kerana dengannya, kita gunakan hasil tambah mengganti hasil darab, dan kuasa eksponen dihapuskan.

$$\begin{aligned} \mathcal{L} &= \ln [\prod_i f(x_i, \theta)] \\ &= \sum_i \ln f(x_i, \theta) \end{aligned}$$

Bagi kes f berbentuk Gaussian,

$$l = \sum \left\{ -\frac{1}{2} \ln 2\pi - \ln \sigma_i - \frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right\}.$$

Parameter bagi taburan yang dipertimbangkan adalah μ_i dan σ_i . Sebutan pertama berbentuk pemalar dan tidak memberi kesan dalam proses pemaksimuman. Kita dapati,

$$\frac{\partial l}{\partial \mu_i} = \frac{2}{2} \frac{(x_i - \mu_i)}{\sigma_i^2} = 0$$

dan

$$\frac{\partial l}{\partial \sigma_i} = -\frac{1}{\sigma_i} + \frac{2}{2} \frac{(x_i - \mu_i)^2}{\sigma_i^2} = 0.$$

Jika $\mu_i = \mu$ untuk semua i dan σ_i semua diketahui, maka

$$\frac{\partial l}{\partial \mu} = \sum \frac{(x_i - \mu)}{\sigma_i^2} = \sum \frac{x_i}{\sigma_i^2} - \sum \frac{\mu}{\sigma_i^2} = 0$$

jika

$$\mu = \hat{\mu}.$$

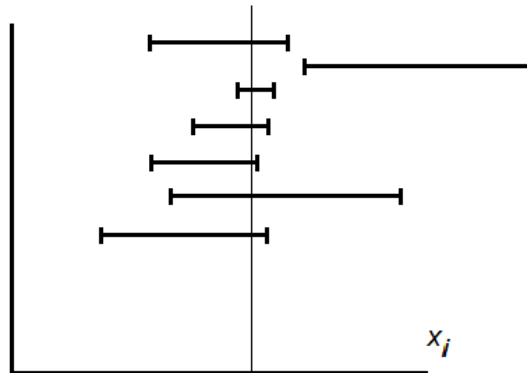
Dengan itu,

$$\hat{\mu} = \frac{\sum x_i / \sigma_i^2}{\sum 1 / \sigma_i^2}.$$

Purata wajar, iaitu purata dengan pemberat, dibuat. Nilai jangkaan bagi anggaran min,

$$E(\hat{\mu}) = \frac{\sum E(x_i) / \sigma_i^2}{\sum 1 / \sigma_i^2} = \frac{\sum \mu / \sigma_i^2}{\sum 1 / \sigma_i^2} = \mu, \text{ yang tidak terpinang. Ukuran dengan } \sigma_i$$

terkecil adalah yang paling penting, yang berikan sumbangan terbesar kepada purata wajar, seperti dalam rajah di bawah.



Varians dalam anggaran bagi min,

$$V(\hat{\mu}) = E(\hat{\mu}^2) - \mu^2 = \left(\frac{1}{\sum 1 / \sigma_i^2} \right)^2 E \left[\left(\sum x_i / \sigma_i^2 \right)^2 \right].$$

Jika x_i bebas, $E(x_i x_j) = E(x_i) E(x_j) = \mu \mu$. Juga, gunakan ungkapan bagi momen kedua, $E[(x - \mu)^2] = E(x^2) - [E(x)]^2$ ataupun $E(x^2) = V(x) + [E(x)]^2$, yakni, $E(x_i^2) = \sigma_i^2 + \mu^2$, kita perolehi,

$$V(\hat{\mu}) = \left(\frac{1}{\sum 1 / \sigma_i^2} \right)^2 \left[\sum \frac{1}{\sigma_i^4} + \mu^2 \sum \frac{1}{\sigma_i^4} + \mu^2 \sum_{i \neq j} \frac{1}{\sigma_i^2 \sigma_j^2} \right].$$

Ini memberikan

$$V(\hat{\mu}) = \left(\frac{1}{\sum 1 / \sigma_i^2} \right)^2 + \mu^2 \left[\frac{\sum 1 / \sigma_i^4 + \sum_{i \neq j} \frac{1}{\sigma_i^2 \sigma_j^2} - \left(\sum 1 / \sigma_i^2 \right)^2}{\left(\sum 1 / \sigma_i^2 \right)^2} \right] = \frac{1}{\sum 1 / \sigma_i^2}.$$

Penganggar ini sebenarnya cekap.

Kalau penukaran pembolehubah dibuat, daripada θ kepada $g(\theta)$,

$$\hat{g} = g(\hat{\theta}),$$

kerana daripada

$$\frac{\partial \mathcal{L}}{\partial g} = \frac{\partial \mathcal{L}}{\partial \theta} \frac{\partial \theta}{\partial g},$$

di maksimum $\frac{\partial \mathcal{L}}{\partial \theta} = 0$, didapati $\frac{\partial \mathcal{L}}{\partial g} = 0$ jika $\frac{\partial \theta}{\partial g} \neq 0$. Dalam kes multivariat, suatu matriks transformasi perlu digunakan.

Beberapa komen dibuat berkenaan kaedah kebolehjadian maksimum:

- Jika saiz sampel besar, kebolehjadian maksimum memberikan anggaran yang bitara, takpincang, varians minimum

- Jika anggaran cekap wujud, ia boleh ditemui oleh kebolehjadian maksimum
- Kaedah kebolehjadian maksimum sesuai bila parameter-parameter taburan berbilangan terhad, tertutup, dan tidak berada di sempadan domain.
- Anggaran kebolehjadian maksimum seringkali yang paling mudah dikira bagi kes-kes kompleks
- Kaedah kebolehjadian maksimum adalah cukup. Terutama bagi sampel-sampel kecil, ia lebih baik daripada kaedah-kaedah yang bergantung kepada membina data
- Kaedah kebolehjadian maksimum tidak semestinya tegap atau tahan lasak
- Kaedah kebolehjadian maksimum tidak memberi jalan untuk menguji kebaikan padanan
- Untuk menggunakan kaedah kebolehjadian maksimum, fungsi taburan kebarangkalian bagi ukuran mesti diketahui

6.4.2 χ^2 – ki kuasadua

Katakan kita ada sampel x_i yang tertabur secara normal atau Gaussian $N(x_i, \mu, \sigma_i)$. Kebolehjadian,

$$\mathcal{L} = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i - \mu)^2}{2\sigma_i^2}}.$$

Jika didefinisikan

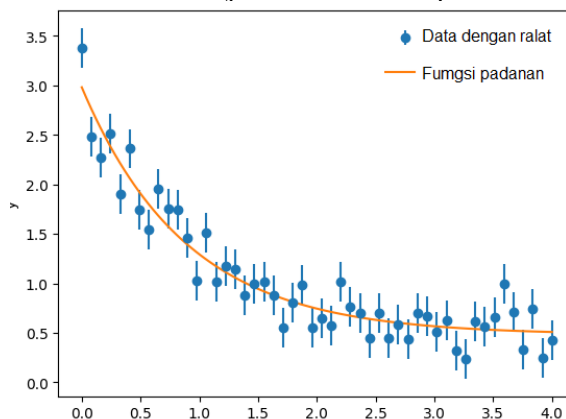
$$d_i = \frac{x_i - \mu}{\sigma_i}$$

maka maksimum $\mathcal{L}$ bermakna minimum $\sum d_i^2 \equiv \chi^2$.

Parameter θ dianggarkan menerusi padanan kepada nilai-nilai yang memberikan χ^2 minimum. Katakan kita tahu fungsi taburan kebarangkalian diberi oleh $f(x_i, \theta)$. Ukuran y_i dibuat pada kedudukan x_i , dengan ralat σ_i . Maka θ yang meminimalkan

$$\chi^2 = \sum_i \frac{(y_i - f(x_i, \theta))^2}{\sigma_i^2}$$

diambil sebagai anggaran kepada parameter-parameter berkenaan. Ia memberikan fungsi yang paling hampir dengan ukuran setelah diwajarkan ralat setiap ukuran. Suatu contoh ditunjukkan di bawah.



Nilai χ^2 selepas padanan memberi ukuran kebaikan padanan. Taburan piawai χ^2 dirujuk untuk mendapatkan kebarangkalian padanan menurut kebarangkalian nilai χ^2 diperolehi.

Kaedah ki kuasadua ini berguna untuk ukuran-ukuran dengan kebarangkalian normal dengan varians berbeza (heteroskedastisiti). Bagi ukuran daripada penghistogramman, taburan Poisson boleh diandaikan, yang hampir serupa dengan taburan normal, dengan nilai varians sama dengan nilai ukuran.

6.4.3 Kuasadua terendah

Bagi ukuran-ukuran dengan ralat yang sama (homoskedastisiti), kita hanya perlu cari minimum hasil tambah kuasadua-kuasadua $\sum_i [y_i - f(x_i, \theta)]^2$, tanpa memerlukan nilai σ_i . Ini ialah kaedah kuasadua terendah.

Kaedah ini biasa digunakan dalam regresi linear dan padanan polinomial. Regresi linear ialah padanan garis lurus kepada data.

Tugasan 12

1. Plot taburan Poisson, $P_{\mu}(n) = \frac{e^{-\mu} \mu^n}{n!}$, dengan digit terakhir nombor matrik anda, + 2.5, digunakan sebagai nilai μ . Sebutkan nilai m ini.

Cari nilai n terendah yang mana $P_{\mu}(n) < 0.01$. Namakan ini n_{maks} .

Jana suatu set nilai-nilai 1000 integer rawak n yang diberikan oleh taburan Poisson ini, menggunakan kaedah Monte Carlo penerimaan-penolakan. Jana dalam selang $[0, n_{\text{maks}}]$.

Gunakan penghistograman untuk menganggar taburan yang dijana (guna bilangan bin yang munasabah).

Guna juga jarak jiran terdekat untuk penganggaran. (Bagi setiap nilai n yang munasabah, plotkan $1/(\text{jarak ke nilai terdekat dalam set})$.)

Kira nilai min dan varians bagi set nilai-nilai ini. Apakah hubungannya dengan μ ?

2. Ukuran-ukuran berikut didapati.

7.117	4.689	1.718	8.062	3.117	3.994	3.821	1.981	3.730	2.817
4.204	1.242	5.017	3.501	2.518	2.692	3.055	2.669	5.196	2.463
0.252	4.779	3.545	3.712	2.908	3.553	1.485	2.301	2.937	5.789
1.108	1.657	2.624	1.697	3.504	6.052	2.383	3.923	4.257	5.337
5.264	3.910	2.783	2.727	1.427	5.250	4.315	3.023	3.556	2.964
2.516	3.108	3.898	2.609	4.216	3.066	3.457	5.214	4.302	5.458
4.628	3.519	3.736	4.986	1.444	5.675	4.146	1.790	3.111	2.544
0.272	4.438	2.673	2.321	2.698	3.504	3.738	3.077	2.880	3.339
2.602	1.980	3.247	3.165	3.721	2.616	3.535	2.519	2.950	3.648
1.400	1.903	1.698	5.148	4.023	1.738	3.841	3.588	3.379	4.131
3.776	1.953	1.293	2.184	2.840	1.431	1.009	1.621	3.868	3.209
2.864	2.552	3.378	2.777	3.002	3.135	2.588	9.685	3.117	2.824
0.497	3.002	2.559	3.490	3.574	5.457	5.040	2.946	3.417	3.302
1.287	2.644	6.086	3.102	2.414	5.277	2.307	4.122	2.727	2.948
3.160	1.937	3.634	5.564	2.865	3.424	5.521	3.102	3.130	2.755
3.342	4.907	5.577	1.925	1.754	4.208	2.047	1.348	3.768	3.668
3.142	2.911	1.624	3.897	2.918	6.032	3.750	7.921	2.717	1.127
5.814	1.662	3.902	2.667	3.276	2.345	4.925	2.487	3.436	2.239
2.509	3.437	2.476	5.606	0.993	4.471	2.212	3.100	3.737	2.474
2.905	3.713	4.550	3.725	6.697	4.383	2.470	6.189	2.993	3.369

Lakukan penghistograman untuk penganggaran tak berparameter kepada data ini.

Plot nilai anggaran y_i di tengah selang bin-bin histogram x_i , dengan ralat statistik di setiap bin sebagai ralatnya ε_i , dikira sebagai $\sqrt{y_i}$ (dianggap statistik Poisson). Plot titik anggaran dengan palang ralat, atas plot baharu.

Katakan kita jangkakan taburan ini diberikan secara teorinya oleh fungsi Breit-Wigner, $f(x) \propto \frac{1}{(x-x_0)^2 + \gamma^2}$.

Lakukan anggaran berparameter untuk memadankan nilai x_0 dan γ , bagi kedudukan puncak dan separuh

lebar di separuh maksimum (HWHM): minimumkan $\chi^2 \equiv \sum_i \frac{(y_i - f(x_i))^2}{\varepsilon_i^2}$ terhadap x_0 dan γ . (Boleh guna

`scipy.optimize.minimize` dalam Python.)

Plot lengkung teori ini di atas plot data+ralat tadi.