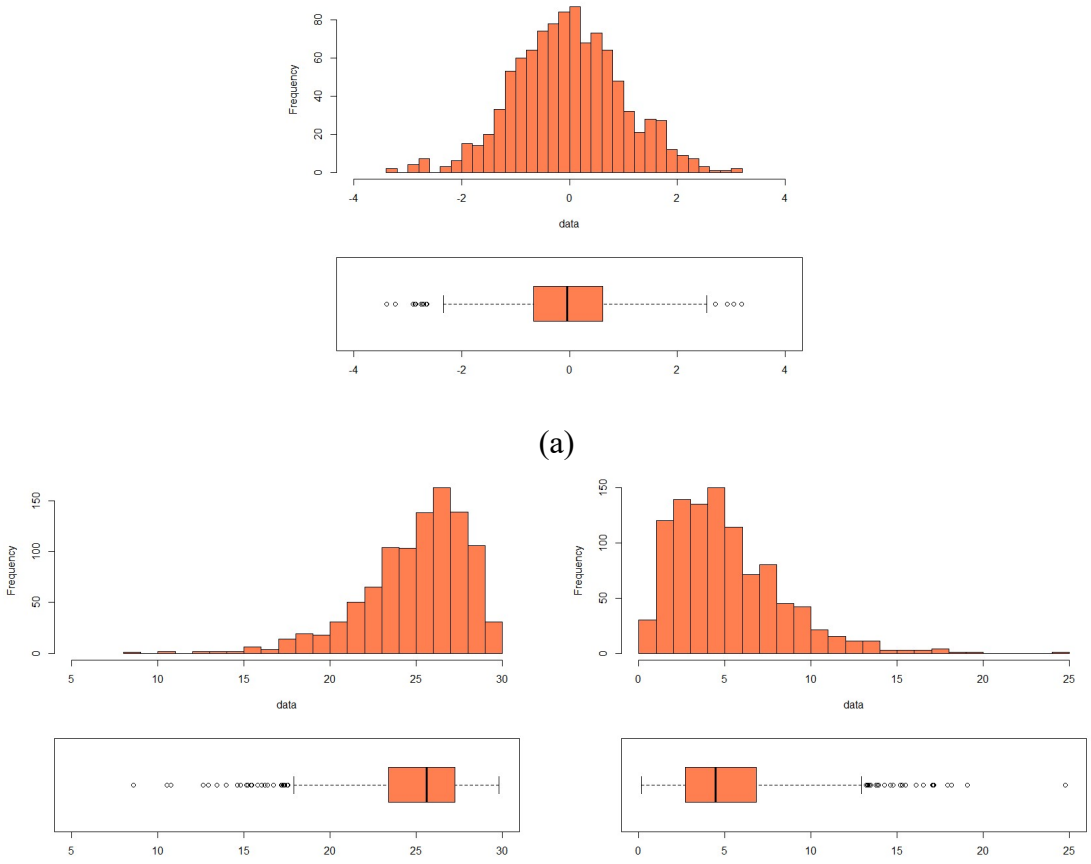


Gambar 1.9 Boxplot yang berpadanan dengan histogram

Sementara itu, pada data dengan sebaran menjulur ke kiri (Gambar 1.9(b)) nampak bahwa bagian gambar kotak dan garisnya akan cenderung lebih panjang pada bagian kiri dibandingkan kotak dan garis yang di bagian kanan. Sebaliknya untuk data dengan sebaran menjulur ke kanan, maka bagian kanan kotak dan garis yang lebih panjang.

1.4. Dugaan Bentuk Fungsi Kepekatan

Histogram dan boxplot, seperti yang telah didiskusikan sebelumnya, dapat dijadian alat untuk mengenali bentuk sebaran dari data yang kita miliki. Pada situasi tertentu kita ingin memperoleh dugaan fungsi sebaran berupa fungsi kepekatan peluang dari data. Fungsi kepekatan merupakan fungsi yang menyatakan struktur peluang dari setiap kemungkinan nilai peubah. Jika kita memiliki fungsi kepekatan dari suatu peubah maka kita dapat menentukan besarnya peluang untuk sembarang selang nilai yang kita inginkan.

Pengetahuan terhadap fungsi kepekatan ini memiliki manfaat pada beberapa hal. Pertama, kita dapat mengenali dengan baik sebaran data jika kita mengetahui fungsi kepekatannya. Seperti yang telah dibicarakan di atas, dengan fungsi kepekatan ini kita dapat membangkitkan nilai peluang untuk berbagai selang nilai. Informasi peluang ini penting dalam kaitannya dengan eksplorasi data secara umum.

Manfaat lain dari pengetahuan terhadap fungsi kepekatan adalah bahwa fungsi ini akan banyak membantu dalam proses analisis dengan teknik simulasi. Berbagai teknik simulasi statistika memerlukan pembangkitan bilangan acak dengan sebaran yang spesifik sesuai dengan karakteristik populasi yang diinginkan. Pengetahuan mengenai fungsi kepekatan ini akan membantu simulasi memperoleh data yang lebih sesuai yang pada ujungnya akan memberikan kesimpulan hasil simulasi secara sah dan memuaskan.

Penduga Naïve

Pendekatan pertama untuk menduga fungsi kepekatan peluang berdasarkan data contoh yang kita miliki adalah yang disebut penduga naïve. Dengan pendekatan ini, penduga bagi fungsi kepekatan dari suatu peubah X adalah

$$\hat{f}_h(x) = \frac{1}{2hn}(\text{banyaknyadata pada selang}(x-h, x+h))$$

dengan n adalah ukuran contoh dan h adalah suatu konstanta yang merupakan setengah dari lebar selang.

Berikut ini adalah program R untuk menghasilkan penduga naïve bagi sebaran data peubah `bwt` (bobot bayi saat lahir) yang ada pada data `lowbwt` yang sudah disinggung sebelumnya. Selain menghitung nilai dugaan fungsi kepekatan $f(x)$, program ini juga menggambarkan secara tumpang tindih histogram data dan fungsi kepekatannya.

PROGRAM R 1.3

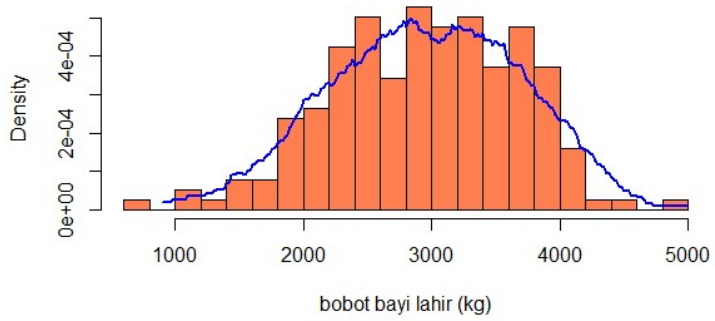
```
data.lahir <- read.csv("lowbwt.csv")
bobot <- data.lahir$bwt
hist(bobot, breaks=25, col="coral",
     xlab="bobot bayi lahir (kg)",
     main="",
     freq=FALSE)

h = 500
n = length(bobot)
x <- seq(900, 5000, by=20)
fx <- NULL
for (i in 1:length(x)){
```

```
  fx[i] = sum(ifelse(abs(bobot - x[i]) < h , 1, 0)) / (2*h*n)
}
lines(x, fx, type="l", col="blue", lwd=2)
```

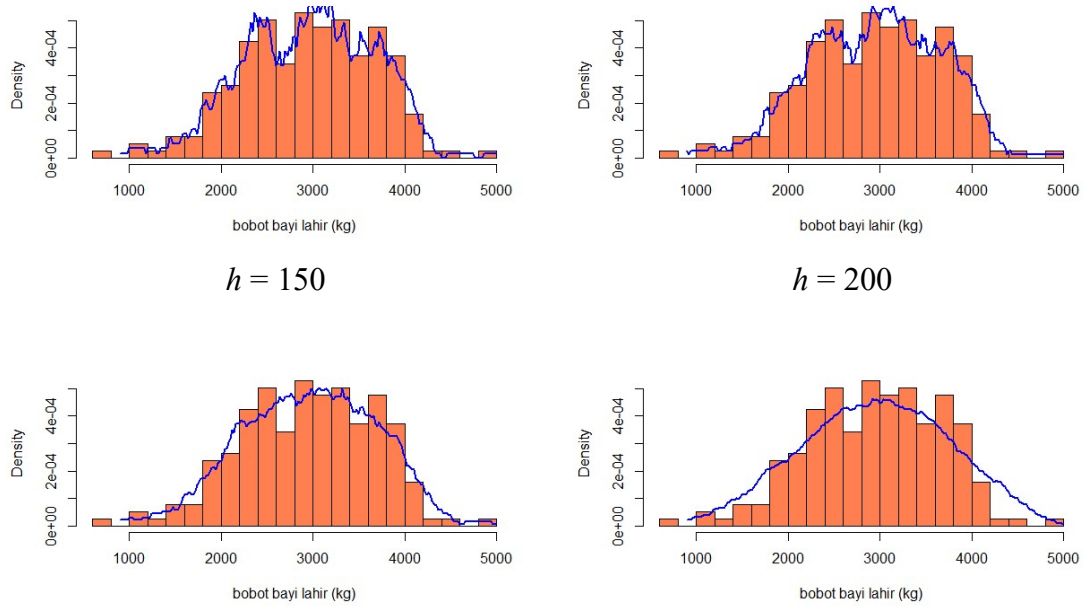
Program menghitung nilai dugaan $f(x)$ dapat dijelaskan sebagai berikut. Perintah `sum(ifelse(abs(bobot - x[i]) < h , 1, 0))` digunakan untuk menghitung banyaknya amatan yang berada pada selang $(x - h, x + h)$ dengan terlebih dahulu mengkonversi amatan yang bernilai dalam selang tersebut menjadi 1 dan yang di luar selang dikonversi menjadi 0. Baris lengkap dari perintah `fx[i]` adalah menyimpan nilai fungsi kepekatan untuk setiap nilai x pada selang antara 900 hingga 5000 kg. Karena tidak mungkin kita menghitung $f(x)$ pada sembarang nilai, pada program ini hanya nilai bobot atau x antara 900 hingga 5000 dengan interval 20 kg saja yang dihitung. Semakin kecil interval itu akan membuat komputasi semakin lama karena semakin banyak nilai x yang diduga nilai $f(x)$-nya.

Dalam program ini besaran h yang digunakan adalah 500, dan grafik yang diperoleh dari program tersebut adalah Gambar 1.10 dimana kurva biru merupakan kurva $f(x)$.



Gambar 1.10 Histogram hasil program R 1.3

Pemilihan nilai h (lebar jendela) akan mempengaruhi bentuk kurva dari dugaan $f(x)$. Kurva $f(x)$ yang diperoleh dari nilai h yang besar akan cenderung menghasilkan kurva yang sangat mulus, sedangkan nilai h yang kecil akan menghasilkan kurva yang bergerigi (*spiky*). Untuk mempermudah menjelaskan fenomeda ini, Gambar 1.11 menyajikan ilustrasi kurva $f(x)$ pada kasus yang sama dengan yang diberikan sebelumnya pada data bobot lahir bayi dengan menggunakan beberapa nilai h yang berbeda.



$h = 150$ $h = 200$

$h = 350$ $h = 750$

Gambar 1.11 Histogram bobot bayi lahir dengan nilai h yang berbeda

Formulasi penduga naïve dari fungsi kepekatan yang telah disampaikan di atas dapat pula ditulis ulang dalam bentuk

$$\hat{f}_h(x) = \frac{1}{hn}\sum_{i=1}^{n} w\left(\frac{x - x_i}{h}\right)$$

dengan x_i adalah nilai data pada amatan ke-i, untuk $i = 1, 2, \ldots, n$, sedangkan $w(\,)$ adalah fungsi yang didefinisikan sebagai

$$w(z) = \begin{cases} 1/2 & \text{jika } |z| < 1 \\ 0 & \text{selainnya} \end{cases}$$

Bentuk di atas akan membantu untuk menuliskan bentuk umum penduga berikutnya yang disebut sebagai penduga kernel.

Penduga Kernel

Penduga kernel merupakan proses pemulusan dari penduga fungsi kepekatan berdasarkan data contoh yang diperoleh menggunakan formula

$$\hat{f}_h(x) = \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Jika pada penduga naïve, nilai bobot $w(\,)$ diberikan sama kepada setiap nilai yang berada dalam selang (x-h, x+h) yaitu sebesar $\frac{1}{2}$, pada penduga ini bobot diberikan sebesar $K(\,)$ yang merupakan fungsi kernel dan nilainya akan secara umum mengecil jika amatannya semakin jauh dari titik pusat selang yaitu x. Fungsi $K(\,)$ harus memenuhi syarat sebagai fungsi yang bersifat non negatif dan integralnya sama dengan 1 (satu).

Beberapa fungsi kernel yang biasa digunakan adalah:

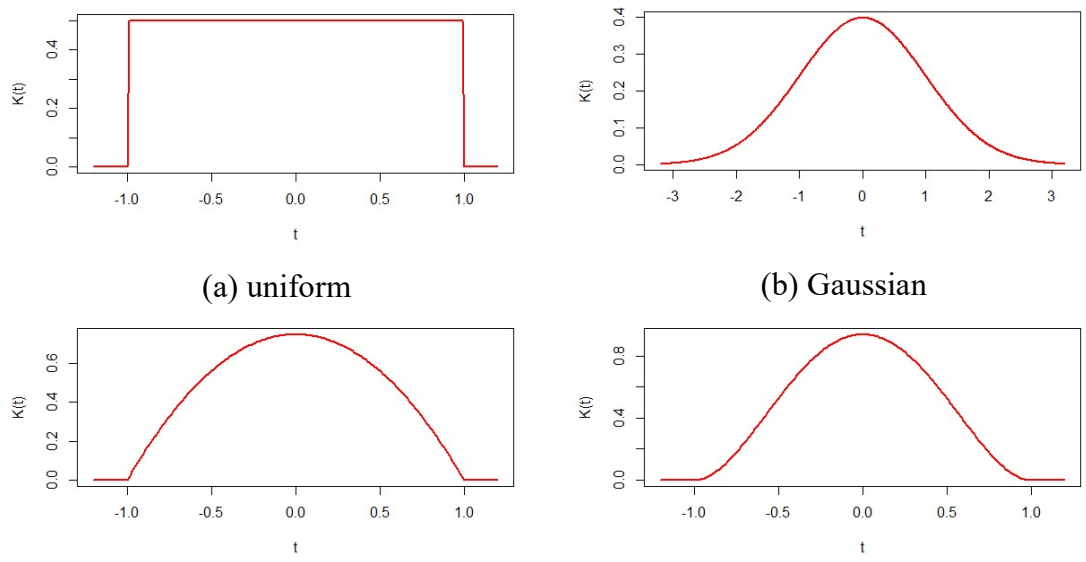
1. Uniform Kernel, $\mathrm{K(t)} = \frac{1}{2}\mathrm{I}(|\,\mathrm{t}\,| \leq 1)$
2. Gaussian Kernel, $K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$
3. Epanechnikov Kernel

$$K(t) = \begin{cases} \frac{3}{4}(1-t^2), & |\,t\,| < 1 \\ 0 & \text{otherwise} \end{cases}$$

4. Biweight Kernel

$$K(t) = \begin{cases} \frac{15}{16}(1-t^2)^2, & |\,t\,| < 1 \\ 0 & \text{otherwise} \end{cases}$$

Gambar 1.12 menampilkan kurva dari keempat fungsi kernel di atas.

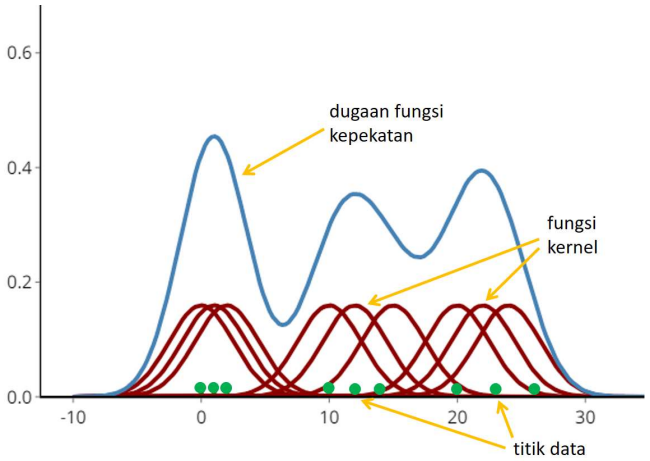


(a) uniform (b) Gaussian

(c) Epanechnikov (d) Biweight

Gambar 1.12. Kurva – kurva fungsi kernel

Pada prinsipnya, penduga kernel untuk fungsi kepekatan tidak lain adalah jumlah dari fungsi kernel. Ilustrasi pada Gambar 1.13 menggambarkan bagaimana penduga kernel bekerja. Di sekitar setiap titik amatan (titik berwarna hijau) dapat diperoleh kurva fungsi kernel. Jika ada beberapa titik amatan yang nilai berdekatan maka akan ada kurva fungsi kernel yang saling berdekatan pula sehingga pada saat dijumlahkan di sekitar nilai tersebut akan memiliki nilai fungsi kepekatan yang lebih tinggi dibandingkan titik-titik lain yang amatannya sedikit.

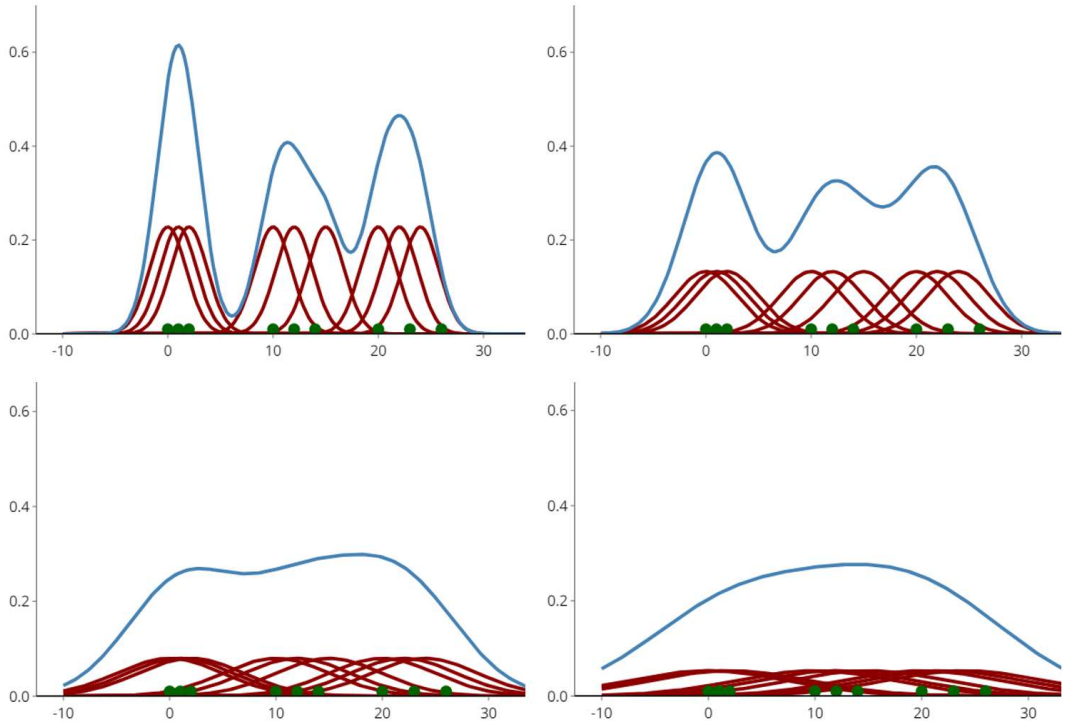


Gambar 1.13 Ilustrasi penduga kernel

Selanjutnya perlu dipahami bahwa bentuk kurva penduga fungsi kepekatan akan sangat tergantung pada lebar jendela (*bandwidth*) h. Jika lebar jendela h kecil maka kurva fungsi kernel akan cenderung lebih kurus dan memberi nilai 0 pada banyak nilai x. Karenanya maka hasil penjumlahan fungsi kernel akan berbentuk lebih bergerigi (*spiky*).

Sementara itu nilai lebar jendela yang besar akan membuat kurva fungsi kernel menjadi melebar dan tidak banyak yang bernilai nol, sehingga ketika dijumlahkan akan memperoleh penduga fungsi kepekatan yang cenderung datar dan tidak banyak lembah.

Gambar 1.14 menyajikan ilustrasi bagaimana bentuk dari fungsi kernel akan berubah-ubah jika lebar jendela diubah dan ini pada akhirnya akan mempengaruhi bentuk dari penduga fungsi kepekatan peluang.



Gambar 1.14 Ilustrasi fungsi kernel dengan perbedaan *bandwith*

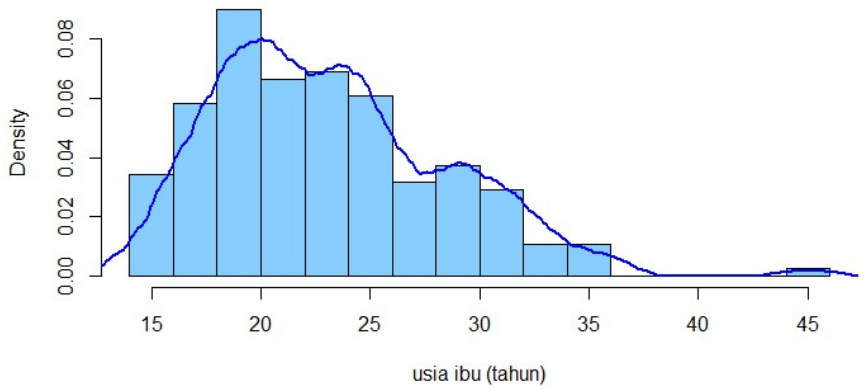
Dari ilustrasi ini jelas bahwa jika lebar jendela dibuat terlalu kecil maka kurva fungsi kepekatan akan cenderung banyak lembah dan puncak, sedangkan kalau terlalu besar akan menghasilkan kurva yang semakin mulus. Karenanya ada juga yang menyebut lebar jendela ini sebagai parameter pemulusan (*smoothing parameter*) pada pendekatan kernel.

Pada program R fungsi yang dapat digunakan untuk menghasilkan penduga fungsi kepekatan menggunakan metode kernel adalah fungsi `density()`. Berikut ini adalah contoh program yang digunakan untuk mendapatkan penduga fungsi kepekatan menggunakan kernel Gaussian, dan selanjutnya meong-*overlay* grafik kurva fungsi kepekatan dan histogram data.

PROGRAM R 1.4

```
kepekatan <- density(data.lahir$age, bw=1,
                     kernel="epanechnikov")
hist(data.lahir$age, freq=FALSE, breaks=15,
     col="skyblue1", main="", xlab="usia ibu (tahun)")
lines(kepekatan, col="blue", lwd=2,
      main="", ylim=c(0, 0.09))
```

Perintah di atas akan menghasilkan grafik pada Gambar 1.15.



Gambar 1.15 Histogram hasil program R 1.4

Opsi utama yang ada pada fungsi `density()` adalah `bw` dan `kernel`. Opsi `bw` digunakan untuk menentukan lebar jendela yang digunakan sedangkan opsi `kernel` digunakan untuk memilih fungsi kernel apa yang digunakan. Fungsi `density()` ini menyediakan beberapa pilihan fungsi kernel yaitu: "gaussian", "epanechnikov", "rectangular", "triangular", dan "biweight".

1.5. Perbandingan Bentuk Sebaran Beberapa Populasi

Jika kita memiliki data yang berasal dari beberapa (sub)populasi, sering menjadi menarik untuk membandingkan karakteristik dari satu populasi dengan populasi lainnya. Salah satu yang bisa kita bandingkan adalah dengan melihat perbandingan bentuk sebarannya.

Untuk membandingkan bentuk sebaran, tidak mudah melakukannya secara visual menggunakan histogram. Beberapa histogram bisa saja kita buat, namun jika ditumpang-tindihkan dalam satu buah gambar akan terlihat berdesakan dan sulit menangkap informasinya. Cara visualisasi yang bisa digunakan adalah dengan membandingkan bentuk dari fungsi kepekatannya. Karena visualnya hanya berupa kurva, maka jika ada beberapa kurva dalam satu gambar masih mudah bagi kita untuk membaca dan memperoleh informasi dari gambar tersebut.

Fungsi `density()` pada R yang sudah kita gunakan sebelumnya dapat menjadi pilihan. Dengan memisahkan gugus data asal menjadi beberapa subset sesuai dengan kelompok yang ada, kita dapat menjalankan fungsi `density()` pada masing-masing kelompok. Namun pada R juga tersedia package lain yaitu `sm` yang menyediakan fungsi dengan nama `sm.density.compare()` yang dapat digunakan untuk menghasilkan fungsi

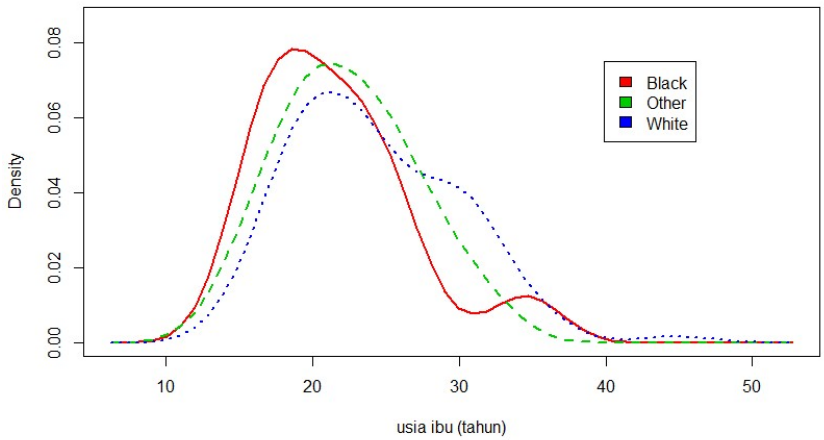
kepekatan dari beberapa kelompok sekaligus dan kemudian menampilkan kurva fungsi kepekatan dalam satu gambar.

Berikut ini adalah ilustrasi program untuk menghasilkan penduga fungsi kepekatan bagi peubah `age` (usia ibu saat melahirkan) pada data `lowbwt` (berupa dataframe dengan nama `data.lahir`) untuk tiga jenis `race` (suku bangsa) ibu.

PROGRAM R 1.5

```
library(sm)
sm.density.compare(data.lahir$age, data.lahir$race,
                   xlab="usia ibu (tahun)",
                   lwd=2)
colfill<-c(2:(2+length(levels(data.lahir$race))))
legend(locator(1), levels(data.lahir$race), fill=colfill)
```

Program di atas akan menghasilkan grafik pada Gambar 1.16. Ada beberapa informasi yang bisa kita peroleh. Pertama, terlihat bahwa puncak dari kurva berwarna merah (sebaran usia ibu melahirkan dari suku bangsa Black) cenderung berada di sebelah kiri dari puncak dua kurva lainnya. Hal ini mengindikasikan bahwa secara rata-rata, ibu-ibu dari ras Black cenderung melahirkan pada usia lebih muda dibandingkan ibu dari ras lain. Informasi lain adalah tentang ukuran penyebaran. Tampak bahwa kurva biru (untuk ras White) cenderung lebih lebar dan lebih pendek daripada dua kurva lainnya. Ini menginformasikan bahwa keragaman usia ibu dari ras White cenderung lebih tinggi dibandingkan dua ras yang lainnya.

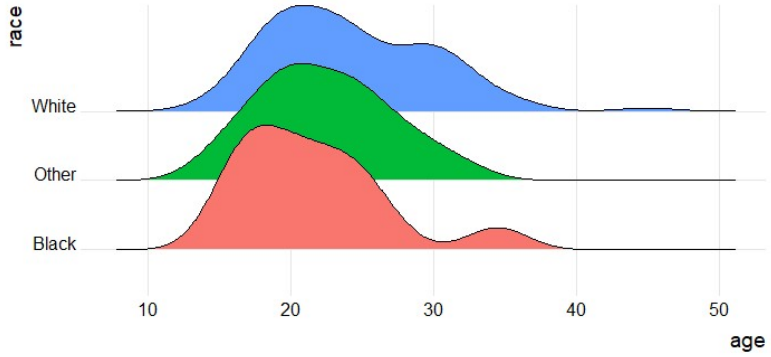


Gambar 1.16 Grafik hasil program R 1.5

Berikut ini dua cara lain di R yang dapat digunakan untuk menghasilkan plot penduga fungsi kepekatan usia ibu dari tiga kelompok ras. Berbeda dengan sebelumnya, tiga plot fungsi kepekatan diletakkan pada sumbu yang disusun paralel, tidak dalam satu sumbu usia, seperti yang diberikan pada Gambar 1.17 dan Gambar 1.18.

PROGRAM R 1.6

```
library(ggridges)
library(ggplot2)
ggplot(data.lahir, aes(x = age, y = race, fill = race)) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")
```

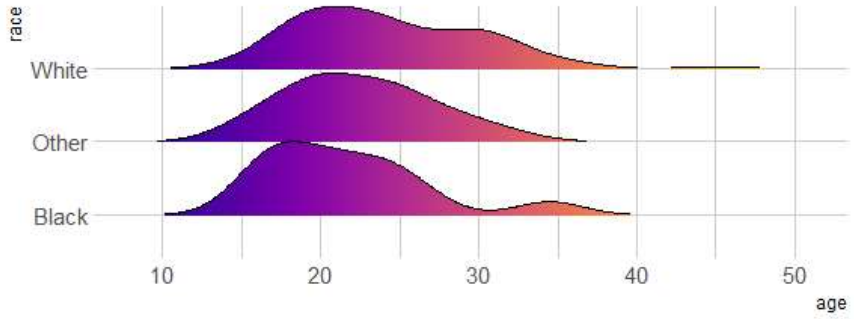


Gambar 1.17 Grafik hasil program R.16

PROGRAM R 1.7

```
library(ggridges)
library(ggplot2)
library(viridis)
library(hrbrthemes)

# Plot
ggplot(data.lahir, aes(x = age, y = race, fill = ..x..)) +
  geom_density_ridges_gradient(scale=1, rel_min_height=0.01) +
  scale_fill_viridis(name = "usia", option = "C") +
  labs(title = '') +
  theme_ipsum() +
  theme(legend.position="none",
    panel.spacing = unit(0.1, "lines"),
    strip.text.x = element_text(size = 8))
```



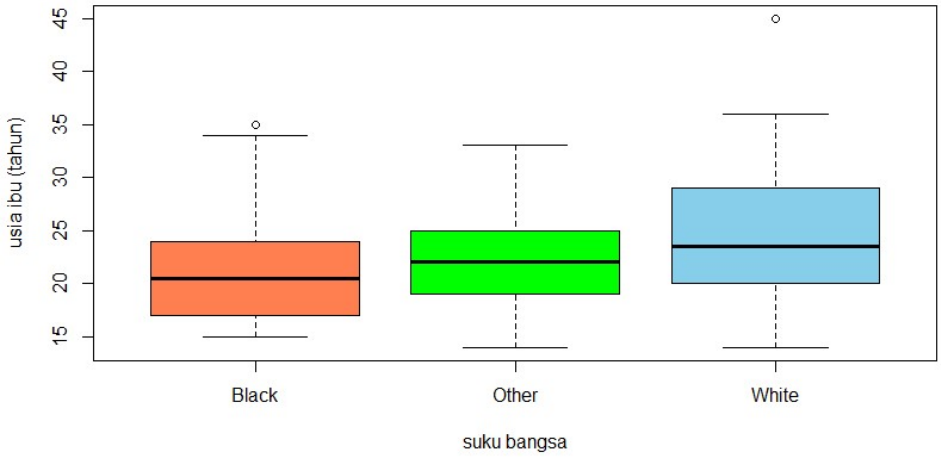
Gambar 1.18 Grafik hasil program R 1.7

Visualisasi lain yang juga dapat digunakan untuk melakukan perbandingan sebaran dari beberapa (sub)populasi adalah dengan boxplot. Program R berikut ini dapat digunakan untuk menghasilkan tiga buah boxplot masing-masing untuk tiga jenis ras ibu dari data yang sama dengan ilustrasi sebelumnya.

PROGRAM R 1.8

```
boxplot(data.lahir$age ~ data.lahir$race,
        col=c("coral", "green", "skyblue"),
        ylab="usia ibu (tahun)",
        xlab="suku bangsa")
```

Dari Gambar 1.19 yang dihasilkan kita bisa melihat bahwa kotak dari boxplot ras Black berada pada posisi lebih rencah dibandingkan dua ras yang lainnya yang menunjukkan bahwa secara rata-rata usia ibu ras Black cenderung lebih muda. Informasi mengenai keragaman usia ibu pada ras White yang cenderung lebih besar yang diperoleh dari kurva fungsi kepekatan juga dapat terlihat disini dengan melihat bahwa kotak berwarna biru cenderung lebih besar ukurannya dibandingkan kotak ras White maupun Other.



Gambar 1.19. Boxplot hasil program R 1.6