3. Identifikasi Pencilan Peubah Tunggal dan Penduga Kekar bagi Parameter Pemusatan dan Penyebaran

3.1. Pengantar

Dalam sebuah gugus data, sangat mungkin terdapat amatan-amatan dengan nilai yang berbeda dari sebagian besar amatan yang lain. Perbedaan tersebut bisa sangat besar sehingga amatan tersebut dikenal sebagai pencilan (*outlier*) dari data. Beberapa orang menyebutnya sebagai data dengan nilai ekstrim (*extreme value*).

Berbedanya nilai suatu amatan dapat terjadi secara natural. Misalnya saja pada kumpulan tanaman yang ditanam bersama-sama dalam suatu petak lahan tertentu, dapat saja ada satu atau dua tanaman yang berbeda kecepatan perkembangannya sehingga jauh lebih tinggi dibandingkan dengan yang lain.

Namun ada kalanya juga keberadaan pencilan dalam suatu gugus data disebabkan adanya kesalahan. Kesalahan dapat saja terjadi pada karena data diinput secara manual oleh operator dalam berbagai bentuk kesalahan sehingga ter-*entry* lebih besar atau lebih kecil dari yang seharusnya. Kesalahan juga dapat terjadi karena data digabungkan dari sumber-sumber berbeda dan digabungkan begitu saja padahal masing-masing sumber memiliki satuan yang berbeda.

Proses perekaman data oleh mesin juga masih memungkinkan mengalami kesalahan sehingga datanya sangat berbeda. Misalnya saja data waktu transaksi di supermarket yang tercatat secara otomatis pada mesin atau komputer kasir. Jika ada kerusakan pada baterai CPU komputer yang berakibat pada matinya jam komputer, maka bisa jadi ada transaksi yang tercatat pada pukul dua dini hari padahal supermarket tidak beroperasi pada waktu tersebut. Kesalahan juga dapat terjadi pada pencatatan oleh mesin berbasis sensor jika ada faktor pengganggu tertentu saat perekaman data.

Yang jelas, sangat mungkin dalam data ada amatan-amatan dengan nilai yang berbeda atau jauh berbeda dari sebagian besar amatan lainnya.

Pengenalan atau pengidentifikasian terhadap pencilan dalam suatu gugus data bisa jadi merupakan proses awal yang harus dilakukan karena beberapa alasan berikut:

1. Amatan pencilan dapat mengganggu proses analisis data. Keberadaan data pencilan dapat berakibat pada diperolehnya nilai-nilai dugaan parameter yang jauh dari yang

seharusnya, misalnya saja pada pendugaan nilai rata-rata dan ragam populasi. Hal demikian juga dapat terjadi pada pemodelan regresi linear dimana penduga koefisien kemiringan garis bisa sangat berbeda dari yang seharusnya didapatkan.

2. Amatan pencilan merupakan informasi penting dan interes dari analisis. Berikut ini beberapa ilustrasi kegiatan pendeteksian pencilan yang penting dilakukan untuk menghasilkan informasi tertentu:
   - *Fraud Detection* dalam Penggunaan Kartu Kredit
     Ketika kartu kredit seseorang dicuri, perilaku belanjanya akan berbeda dengan perilaku yang biasa dilakukan oleh pemilik aslinya. Pengenalan pola belanja yang tidak wajar akan membantu mengenali penyalahgunaan kartu ini. Pendeteksian pola yang berbeda dari biasanya dapat digunakan untuk meminimalkan resiko kerugian yang timbul akibat *fraud* tersebut.
   - Kesehatan
     Gejala yang tidak biasa atau hasil tes laboratorium yang sangat berbeda dapat menjadi indikasi adalah permasalahan kesehatan tertentu dari seorang pasien. Pemeriksaan terhadap keberadaan hasil tes yang jauh dari kondisi yang biasanya dapat membantu secara cepat mengetahui permasalahan yang ada.
   - Produk cacat
     Pemeriksaan pencilan juga berguna dalam proses pengendalian kualitas produk. Adanya karakteristik produk yang berbeda (diindikasikan keluar dari batas kendali, *out-of-control*) menjadi petunjuk bagi industri untuk melakukan pemeriksaan terhadap proses yang berlangsung. Dapat saja pencilan tersebut muncul karena ada masalah dalam mesin yang digunakan, perubahan pada kualitas bahan baku, kesalahan oleh operator, dan sebab-sebab lainnya.
   - Hotspot Spasial
     Hotspot didefinisikan sebagai titik pada peta spasial yang memiliki nilai jauh lebih besar dari nilai-nilai yang ada pada titik di sekitarnya. Pengidentifikasian lokasi hotspot penting baik dalam bidang kesehatan, kemiskinan, atau yang lainnya sehingga dapat ditentukan prioritas penyelesaian permasalahan.

Diskusi mengenai keberadaan amatan pencilan dapat ditemui untuk kasus peubah tunggal dan peubah ganda. Pada bab ini sementara kita akan diskusikan hanya kasus pencilan pada peubah tunggal.

3.2. Teknik Identifikasi Pencilan Peubah Tunggal

Pada pembahasan pencilan peubah tunggal, yang disebut sebagai pencilan adalah amatan yang memiliki nilai berbeda (jauh lebih besar atau jauh lebih kecil) dari sebagian besar amatan yang lain. Karenanya intuisi paling sederhana yang bisa dikemukakan untuk mengidentifikasi pencilan peubah tunggal adalah jika amatan tersebut jauh dari nilai ukuran pemusatan (misalnya rata-rata).

Berdasarkan logika tersebut, dengan mengasumsikan data cenderung memiliki sebaran yang mendekati normal, maka berapa simpangan baku dari rata-rata jarak suatu amatan dapat digunakan sebagai batasan bagi amatan untuk dikategorikan sebagai pencilan. Konstanta pengali yang banyak digunakan adalah 3. Suatu amatan dikatakan sebagai pencilan jika selisih antara nilai amatan tersebut dengan rata-rata ($\bar{x}$) lebih besar dari $3s$ dengan s adalah nilai simpangan baku. Dengan kata lain, suatu amatan akan disebut sebagai pencilan jika berada di luar selang

$$(\bar{x} - 3s, \bar{x} + 3s).$$

PROGRAM R 3.1

```
#membaca data
data.lahir <- read.csv("lowbwt.csv")
bobot <- data.lahir$bwt
m <- mean(bobot)
s <- sd(bobot)
pencilan <- (bobot > m+3*s) | (bobot < m-3*s)

#menghitung banyaknya amatan pencilan
sum(pencilan)

#mengidentifikasi nomor amatan yang menjadi pencilan
which(pencilan)

#nilai pencilan
bobot[which(pencilan)]
```

```
> sum(pencilan)
[1] 1
> which(pencilan)
[1] 1
> bobot[which(pencilan)]
[1] 709
```

Pendekatan lain yang bisa dikerjakan adalah dengan menggunakan batasan lain. Batas yang digunakan untuk mengidentifikasi pencilan adalah batas bawah dan batas atas dengan formula:

- Batas Bawah = Q1 – 1.5 * IQR
- Batas Atas = Q3 + 1.5 * IQR

dengan IQR adalah *Inter Quartile Range* atau Jarak Antar Kuartil yang merupakan selisih antara Q_3 dengan Q_1. Amatan yang bernilai lebih kecil daripada Batas Bawah atau amatan yang bernilai lebih besar daripada Batas Atas diidentifikasi sebagai pencilan.

PROGRAM R 3.2

```
#identifikasi pencilan
nilai.pencilan <- boxplot.stats(bobot)$out
which(bobot == nilai.pencilan)
nilai.pencilan
```

```
> which(bobot == nilai.pencilan)
[1] 1
> nilai.pencilan
[1] 709
```

3.3. Penduga Kekar bagi Rata-Rata

Keberadaan pencilan dalam suatu gugus data akan berpengaruh terhadap penduga ukuran pemusatan yaitu rata-rata. Adanya data ekstrim yang besar akan menyebabkan rata-rata contoh menjadi lebih besar dari yang seharusnya, sedangkan data ekstrim yang kecil akan menyebabkan nilai rata-rata contoh menjadi lebih rendah dari yang semestinya. Karenanya kita menyebutkan bahwa rata-rata tidak bersifat kekar (*robust*) karena nilainya mudah terganggu oleh keberadaan pencilan terutama jika ukuran gugus datanya tidak besar.

Untuk itu, diperkenalkan beberapa penduga rata-rata populasi yang bersifat kekar dan tidak banyak terpengaruh oleh keberadaan pencilan. Beberapa penduga kekar tersebut akan didisksikan berikut ini.

1. Trimmed Mean

Rataan terpangkas (*trimmed mean*) merupakan rata-rata dari data yang ada di bagian tengah data, tepatnya data di $1 - 2\alpha$ bagian tengah data dengan $0 < \alpha < 1$. Dengan kata lain, sebelum menghitung rata-rata kita seolah-oleh menyisihkan $[\alpha n]$ amatan terbesar serta $[\alpha n]$ amatan terkecil dengan $[c]$ menyatakan fungsi bilangan bulat terdekat dari c.

Rataan terpangkas 2α selanjutnya dapat dituliskan sebagai

$$\bar{y}_{T\alpha} = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} y_{(i)}$$

dengan $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ adalah data terurut dari contoh $y_1, y_2, \ldots, y_n$.

Sebagai contoh andaikan kita memiliki contoh yang berisi 20 amatan (yang sudah terurut) sebagai berikut

12, 12, 12, 13, 13, 14, 16, 17, 18, 18, 19, 19, 19, 19, 20, 20, 21, 22, 22, 67

Rataan dari contoh di atas adalah 19.65 dan cenderung tidak menggambarkan pemusatan data secara baik karena memiliki nilai yang lebih besar dari sebagian besar amatan. Hal ini tentu saja dikarenakan ada satu amatan yang ekstrim lebih besar dibandingkan nilai dari amatan-amatan lainnya.

Rataan terpangkas 10% dari gugus data tersebut diperoleh dengan membuang [5% × 20] = 1 amatan paling besar dan [5% × 20] = 1 amatan paling kecil sehingga tersisa hanya 18 amatan saja. Jika ini dilakukan maka akan diperoleh hasil yang lebih sesuai yaitu 17.44.

Berikut ini adalah ilustrasi penggunaan program R dalam menghitung *trimmed mean* di R menggunakan fungsi `mean()` dengan menambahkan opsi `trim`. Penentuan opsi `trim = 0.05` mengindikasikan bahwa kita ingin memperoleh rata-rata terpangkas 5% pada masing-masing ujung kiri dan kanan data.

PROGRAM R 3.3

```
> contoh <- c(12, 12, 12, 13, 13, 14, 16, 17, 18, 18,
              19, 19, 19, 19, 20, 20, 21, 22, 22, 67)
> mean(contoh)
 [1] 19.65
> mean(contoh, trim=0.05)
 [1] 17.44444
```

2. Winsorized Mean

Winsorized Mean didefinisikan sebagai

$$\bar{y}_{T\alpha} = \frac{1}{n}\left(([n\alpha]+1)y_{([n\alpha]+1)} + \sum_{i=[n\alpha]+1}^{n-[n\alpha]} y_{(i)} + ([n\alpha]+1)y_{(n-[n\alpha])} \right)$$

Nilai *winsorized mean* diperoleh dengan menghitung rata-rata setelah kita lakukan penggantian nilai terhadap amatan-amatan terbesar dan terkecil. Sebanyak $n\alpha$ amatan terkecil diganti nilainya dengan nilai amatan $y_{([n\alpha]\ +\ 1)}$ dan sebanyak $n\alpha$ amatan terbesar diganti nilainya dengan nilai amatan $y_{(n\ -\ [n\alpha])}$.

Untuk melakukan penghitungan *winsorized mean* di R kita dapat menuliskan program yang menghasilkan fungsi sesuai dengan prosedur yang telah diuraikan sebelumnya. Berikut ini contoh pembuatan fungsi dengan nama `winsorMEAN()` dengan argumen berupa vektor data dan dua buah proporsi data terkecil dan data terbesar yang akan diganti. Default yang diberikan untuk nilai proporsi masing-masing adalah 0.05 dan 0.95 yang mengindikasikan bahwa 5% data terkecil dan 5% data terbesar nantinya akan diganti dengan nilai pada titik tersebut, yaitu $Q(0.05)$ dan Q(0.95).

Jika fungsi sudah dijalankan maka selanjutnya untuk memanggil fungsi tersebut tinggal kita sebutkan `winsorMEAN()` yang diisi dengan argumen data yang mau dicari rata-ratanya dan nilai proporsi yang ingin digunakan. Karena *default* nilai peluang adalah 5% dan 95% maka untuk α = 5% kita bisa saja tidak menyebutkan kembali. Namun untuk proporsi selainnya, misalkan 10% kita perlu menyebutkan seperti yang ada pada program di bawah ini.

PROGRAM R 3.4

```
winsorMEAN<-function(x,probs=c(0.05,0.95)) {
xq<-quantile(x,probs=probs)
x[x < xq[1]]<-xq[1]
x[x > xq[2]]<-xq[2]
return(mean(x))
}

contoh <- c(12, 12, 12, 13, 13, 14, 16, 17, 18, 18, 19, 19, 19, 19, 20,
20, 21, 22, 22, 67)
wm05 <- winsorMEAN(contoh)
wm10 <- winsorMEAN(contoh, probs=c(0.10, 0.90))

wm05
wm10
```

Hasil dari program di atas adalah dua buah nilai winsorized mean masing-masing untuk α sebesar 5% dan 10% yang nilainya adalah sebagai berikut.

```
> wm05
[1] 17.5125
> wm10
[1] 17.4
```

Untuk diketahui, pada `package psych` terdapat fungsi dengan nama `winsor.mean()` yang dapat digunakan menghitung *winsorized mean* dari suatu

vektor data. Berikut ini program yang mengilustrasikan penggunaan fungsi tersebut. Nilai α ditentukan dengan menyebutkannya pada opsi `trim`.

PROGRAM R 3.5

```
contoh <- c(12, 12, 12, 13, 13, 14, 16, 17, 18, 18, 19, 19, 19, 19, 20,
20, 21, 22, 22, 67)

library(psych)
winsor.mean(contoh, trim=0.1)
winsor.mean(contoh, trim=0.05)
```

Hasil dari program di atas adalah sebagai berikut yang nilainya persis sama dengan yang kita peroleh menggunakan fungsi yang kita susun sendiri.

```
> winsor.mean(contoh, trim=0.1)
[1] 17.4
> winsor.mean(contoh, trim=0.05)
[1] 17.5125
```

3. *M-estimators* untuk Rata-Rata

Sebelum mendiskusikan penduga-*M*, kita beberapa hal dasar berikut. Perhatikan seandainya kita memiliki data contoh berukuran n, yaitu $x_1, x_2, \ldots, x_n$ dan kita ingin mendapatkan μ yang meminimumkan

$$\sum_{i=1}^{n}(x_i - \mu)^2$$

Solusi dari proses minimisasi ini adalah $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$.

Secara umum, kita bisa mencari penduga rata-rata μ dengan cara meminimumkan besaran

$$\sum_{i=1}^{n}\rho(x_i - \mu)$$

dengan $\rho(\text{-})$ merupakan suatu fungsi tertentu. Untuk meminimumkan formula di atas, menggunakan turunannya kita dapat memperoleh solusinya dengan menyelesaikan persamaan

$$\sum_{i=1}^{n}\rho'(x_i - \mu) = 0.$$

dimana $\rho'(\text{-})$ merupakan turunan pertama dari fungsi $\rho(\text{-})$.

Penduga M bagi μ merupakan penduga yang didapatkan dengan mencari penyelesaian bagi persamaan

$$\sum_{i=1}^{n} \Psi(x_i - \mu) = 0$$

untuk suatu fungsi Ψ tertentu. Beberapa bentuk fungsi Ψ ini antara lain adalah:

1. Huber, yang memiliki bentuk fungsi

$$\Psi(x) = \begin{cases} -c, & \text{untuk } x < c \\ x, & \text{untuk } |x| \le c \\ c, & \text{untuk } x > c \end{cases}$$

2. Tukey's Bisquare, yang memiliki bentuk fungsi

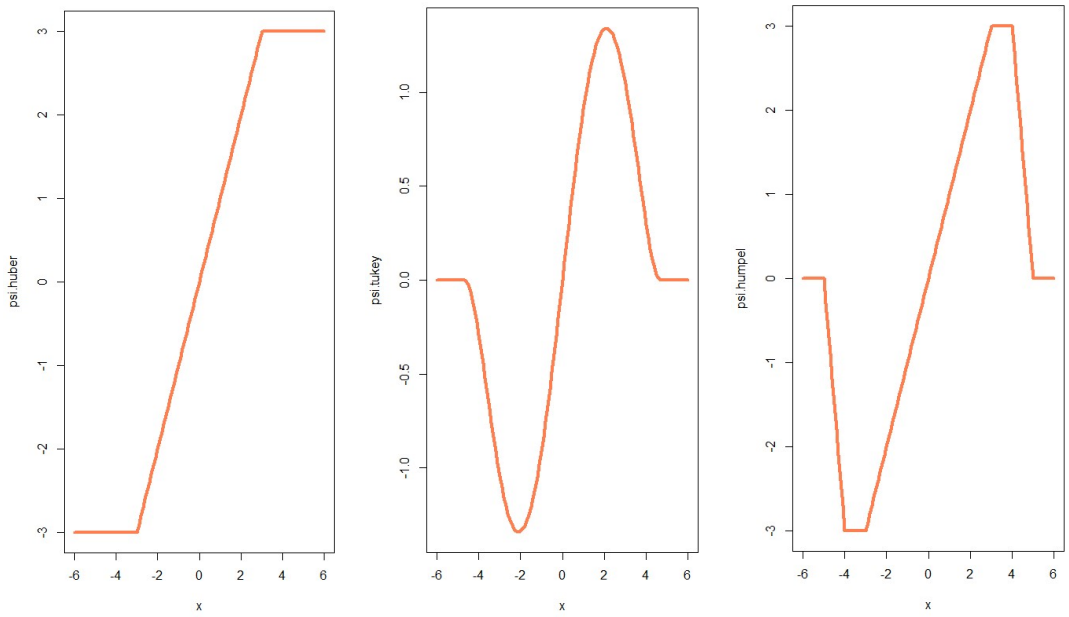
$$\Psi(x) = x\left[1 - \left(\frac{x}{R}\right)^2\right]_+^2$$

dengan $[u]_+ = max\{u, 0\}$. Untuk data dengan sebaran normal, R yang efisien bernilai 4.685.

3. Hampel's Ψ function, yang memiliki bentuk fungsi

$$\Psi(x) = \begin{cases} x, & \text{untuk } |x| \le a \\ a \times sign(x), & \text{untuk } a < |x| < b \\ a\left(\frac{c - |x|}{c - b}\right) \times sign(x), & \text{untuk } b < |x| \le c \\ 0 & \text{untuk } |x| > c \end{cases}$$

Tampilan visual dari ketiga fungsi di atas adalah seperti yang disajikan pada Gambar 3.1.



Gambar 3.1 Tampilan visual beberapa 3 fungsi Ψ untuk *M-Estimators*

3.4. Penduga Kekar bagi Simpangan Baku

Simpangan baku dan juga ragam, merupakan ukuran penyebaran data yang paling populer digunakan dan disajikan oleh peneliti dalam laporan maupun tulisan ilmiah. Namun seperti halnya rata-rata, nilai simpangan baku ini mudah sekali terpengaruh oleh keberadaan pencilan. Jika ada sebuah amatan ekstrim baik yang bernilai jauh lebih besar maupun jauh lebih kecil dibandingkan amatan lainnya, maka simpangan baku akan membengkak menjadi lebih besar dari yang semestinya jika amatan pencilan tersebut tidak ada. Karena itu, menggunakan simpangan baku pada saat data memuat pencilan akan menyebabkan kita salah menafsirkan kondisi data. Kita akan cenderung menganggap data memiliki variasi yang besar padahal sesungguhnya tidaklah persis demikian.

Penduga kekar bagi simpangan baku diperlukan untuk memberikan informasi mengenai simpangan baku yang sesungguhnya tanpa banyak terganggu oleh keberadaan amatan pencilan pada gugus data. Berikut ini diberikan beberapa alternatif penduga yang bersifat kekar untuk menggambarkan ukuran penyebaran data.

1. **Jarak Antar Kuartil (*Inter Quartile Range*/IQR)**
   Jarak antar kuartil (IQR) merupakan selisih antara nilai kuartil ketiga dari data dengan kuartil pertamanya, atau kita tuliskan

$$IQR = Q3 - Q1$$

Pada data yang mengikuti sebaran normal, simpangan baku dapat diperoleh dengan membagi nilai jarak antar kuartil dengan 1.34898. Mengapa demikian?

Secara umum, jika X adalah peubah yang berdistribusi normal maka X dapat dituliskan sebagai $X = mu + sigma\, Z$ dengan mu dan $sigma$ adalah nilai rata-rata dan simpangan baku dari X, sedangkan Z adalah peubah acak normal baku (peubah acak normal dengan rata-rata 0 dan ragam 1). Dengan asumsi sebarannya normal, IQR dari peubah X adalah sama dengan sigma kali IQR dari Z atau dituliskan

$$IQR(X) = sigma * IQR(Z) = sigma * (Q3(Z) - Q1(Z)).$$

Atau dengan kata lain kita bisa menuliskan sebagai

$$sigma = \frac{IQR(X)}{IQR(Z)}$$

dan bisa diverifikasi bahwa $IQR(Z)$ adalah 1.34898 yang menggunakan R bisa dihitung menggunakan perintah `qnorm(0.75)-qnorm(0.25)`.

2. MAD (Median Absolute Deviation)

MAD merupakan penduga kekar paling populer untuk ukuran penyebaran. Nilai MAD didefinisikan sebagai

$$\text{MAD} = \text{median}_i \{ y_j - \text{median}_j\{y_j\}\}$$

dimana $\text{median}_j\{y_j\}$ merupakan median dari n amatan pada gugus data sedangkan median_i menghitung median dari nilai mutlak selisih antara setiap data dengan mediannya.

Pada data yang menyebar normal, simpangan baku dapat diduga sebagai $1.4826 \times MAD$. Alasan mengenai hal tersebut adalah berikut ini.

Jika X adalah peubah yang sebarannya normal, maka $MAD(X) = sigma * MAD(Z)$ dengan *sigma* adalah simpangan baku peubah X sedangkan Z adalah peubah normal baku. Sehingga jelas bahwa

$$sigma = \frac{MAD(X)}{MAD(Z)}$$

dan secara teori dapat ditunjukkan bahwa $MAD(Z)$ = 1/1.4826.

Program berikut menyajikan berapa besaran nilai $MAD(Z)$ dengan cara simulasi. Proses yang dikerjakan adalah membangkitkan contoh acak dari sebaran normal baku atau Z yang berukuran sangat besar (dalam program ini digunakan n =

1.000.000). Kemudian dari contoh yang diperoleh selanjutnya dihiung $MAD(Z)$ dan ditampilkan $1/MAD(Z)$ yang secara teori adalah sebesar 1.4826. Proses tersebut diulang sebanyak 10 kali dan terlihat bahwa kita memperoleh nilai yang dekat dengan nilai 1.4826.

PROGRAM R 3.6

```
for(i in 1:10) {
data <- rnorm(1000000)
print(1/median(abs(data-median(data))))
}
```

```
[1] 1.483130
[1] 1.478438
[1] 1.484341
[1] 1.482961
[1] 1.481436
[1] 1.483619
[1] 1.484163
[1] 1.482785
[1] 1.479327
[1] 1.482337
```

Program R yang dapat digunakan untuk menghitung MAD dari suatu data adalah sebagai berikut. Nilai MAD dari suatu gugus data diperoleh menggunakan fungsi `mad()` dengan menggunakan opsi `constant = 1`. Dalam program ini juga dihitung nilai simpangan baku biasa dan simpangan baku kekar.

PROGRAM R 3.7

```
contoh <- c(12, 12, 12, 13, 13, 14, 16, 17, 18, 18, 19, 19, 19, 19, 20,
20, 21, 22, 22, 67)
nilai.mad <- mad(contoh, constant=1)
simp.baku <- sd(contoh)
simp.baku.kekar <- mad(contoh)
c(nilai.mad, simp.baku, simp.baku.kekar)
```

Nilai-nilai yang dihasilkan dari program di atas adalah berikut ini. Terlihat bahwa nilai simpangan baku data adalah 11.6496 yang sangat besar karena adanya amatan pencilan pada data. Dengan menggunakan pendekatan kekar menggunakan bantuan MAD, diperoleh bahwa penduga simpangan baku adalah sebesar 3.7065.

```
> c(nilai.mad, simp.baku, simp.baku.kekar)
[1]  2.5000 11.6496  3.7065
```

Nilai simpangan baku kekar sebesar 3.7065 diperoleh dari $1.4826 * MAD$ atau $1.4826 * 2.5$.

3. Gini's mean difference

Gini's mean difference didefinisikan sebagai rata-rata selisih nilai antara satu amatan dengan amatan lainnya. Semakin besar nilai *Gini's mean difference* ini tentu saja menjadi indikasi bahwa antar amatan memiliki nilai yang semakin berbeda atau dengan kata lain gugus data kita berisi amatan yang heterogen.

Dalam perhitungannya dilakukan dibuat pasangan-pasangan amatan dan dari setiap pasangan dihitung nilai mutlak selisihnya. Selanjutnya *Gini's mean difference* adalah rata-rata dari selisih tersebut, atau dalam bentuk formula dapat dituliskan sebagai:

$$G = \frac{1}{\binom{n}{2}} \sum_{i<j} |y_i - y_j|$$

Jika gugus data memiliki sebaran normal maka simpangan baku akan sama dengan $\sqrt{\pi} G / 2$, sehingga nilai ini dapat digunakan sebagai ukuran simpangan baku yang robust.

Pada program R, nilai statistik ini dapat dihitung menggunakan fungsi `gini.mean.diff()` yang ada pada `package lmomco` yang penggunaannya dapat dilihat pada ilustrasi program berikut.

PROGRAM R 3.8

```
contoh <- c(12, 12, 12, 13, 13, 14, 16, 17, 18, 18, 19, 19, 19, 19, 20,
20, 21, 22, 22, 67)
library(lmomco)
gini.mean.diff(contoh)$gini
```

```
> gini.mean.diff(contoh)$gini
[1] 8.605263
```