# STK573
# METODE GRAFIK UNTUK ANALISIS DAN PENYAJIAN DATA

## Pendugaan Fungsi Kepekatan Nonparametrik

DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN IPA

INSTITUT PERTANIAN BOGOR

SEMESTER GENAP 2016

# PENDAHULUAN

- Statistics: collection, summarization, presentation, and interpretation of data
- Data are the key to make inferences
- No assumptions about the underlying process that generated these data
- It is assumed parametric model (such as Gaussian with $\mu$ and $\sigma^2$) or nonparametric
- If the **assumed model** is not the correct one, then inferences can be worse and misleading interpretations of the data

# PENDAHULUAN
## WHY NONPARAMETRIC ?

- Parametric:

  strict assumptions that are often violated by real data

  strict hypotheses → if correct, accurate and precise estimates, otherwise very misleading

  linear relationships between the dependent variable and predictor variables (normality, and linearity)

- Nonparametric:

  less-strict assumptions that are less-frequently violated by data

  less conditions → free estimates from hypotheses

  wide range of relationships between the dependent variable and predictor variables (linear, moderately nonlinear, or highly-nonlinear)

# PENDAHULUAN
## NONPARAMETRIC (SMOOTHING)

- a bridge between making no assumptions on formal structure (a purely nonparametric approach) and making very strong assumptions (a parametric approach)
- to identify potentially unexpected structure to more complicated data analysis problems
- to extract more information from the data than is possible purely nonparametrically, as long as the (weak) assumption of smoothness is reasonable
- to provide analyses flexible and robust

An aim of nonparametric techniques is to reduce possible modeling biases of parametric models

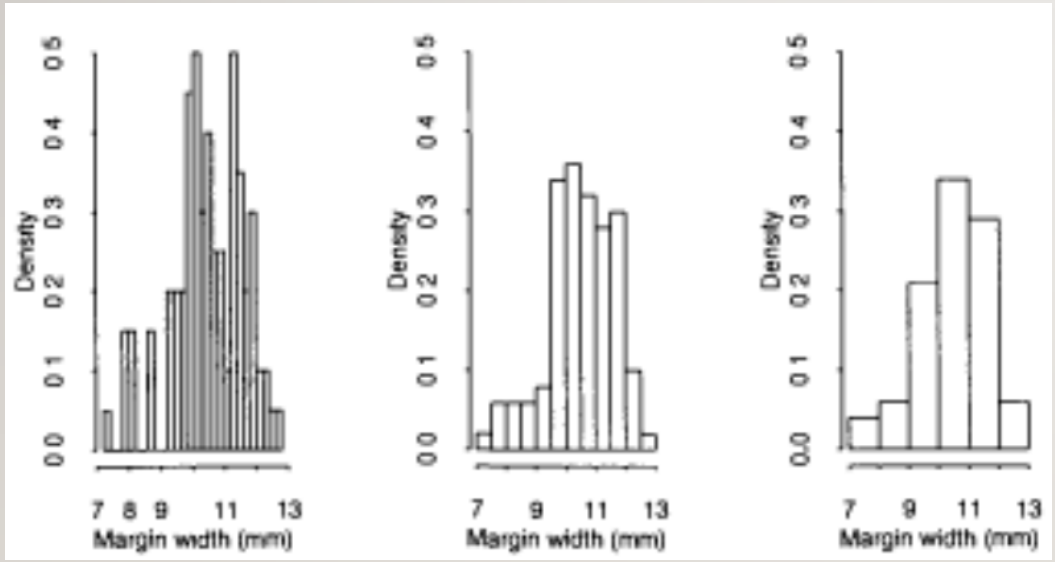
# PENDAHULUAN

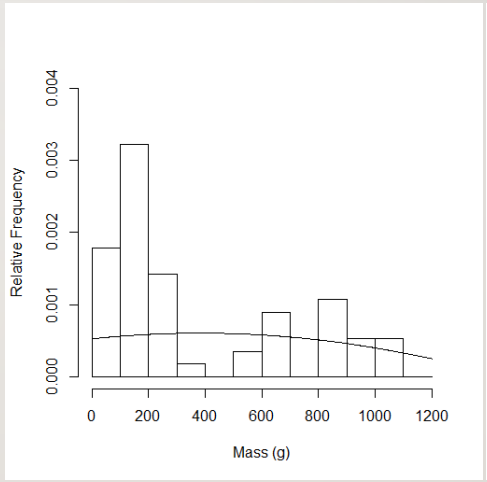
## HISTOGRAM DAN KERNEL

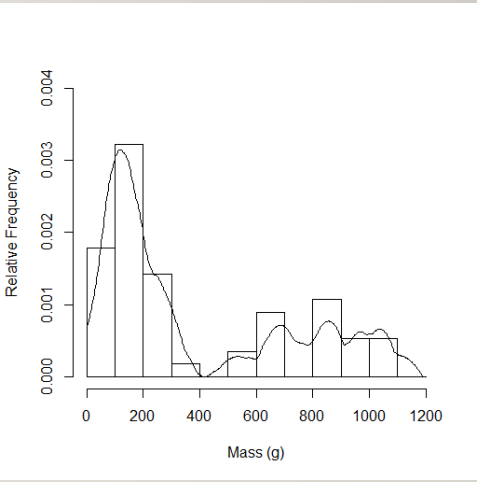
- Penduga suatu fungsi kepekatan dapat dilakukan dalam bentuk histogram dan/atau berdasarkan suatu fungsi kernel atau spline, termasuk prosedur nonparametrik.
- Histogram memberikan gambaran kepekatan secara kasar.
- Kernel atau spline menghasilkan kepekatan yang lebih halus dan mulus (*smooth*).
- Bentuk kepekatan keduanya serupa.
- Tingkat kekasaran histogram tergantung pada **jumlah kelas** sedangkan tingkat kemulusan hasil kernel tergantung pada **lebar jendela** (bandwidth).

- # PENDAHULUAN

## HISTOGRAM DAN KERNEL



# METODE HISTOGRAM

- Deskripsi tentang penyebaran, kemiringan atau kemenjuluran, dan kemungkinan adanya modus ganda
  → **Histogram**

- Gambaran perilaku data sebagai komponen penting dalam analisis data
- Pola data ideal yang simetrik tidak selalu tergambarkan secara baik
  → **Metode Pendugaan Nonparametrik**
  → **Pemulusan**

# METODE HISTOGRAM

- Penyajian data peubah kontinu tunggal dalam bentuk **histogram** sering dan banyak digunakan terutama untuk mengetahui bentuk sebaran data.
- Histogram dibentuk tanpa asumsi suatu model statistik dan tanpa pendugaan parameter-parameternya berdasarkan data sehingga penyajian data dalam bentuk histogram termasuk ke dalam kategori nonparametrik.
- Data dibagi sesuai dengan jumlah kelas yang telah ditetapkan lebih dulu.
- Rentang atau lebar (*interval*) setiap kelas sama yang ditentukan berdasarkan wilayah (*range*) data dibagi dengan banyaknya kelas.
- Setelah titik awal dan akhir setiap kelas ditentukan, sejumlah pengamatan (frekuensi) pada setiap lebar kelas ditentukan dan digambarkan dalam bentuk persegi panjang, sehingga terbentuk **histogram frekuensi**.

# HISTOGRAM

- Histogram merupakan penduga fungsi kepekatan nonparametrik
- Proses penyusunan histogram:
  - Penentuan jumlah kelas (segmen) nilai
  - Penentuan lebar kelas
  - Penentuan lokasi nilai tengah masing-masing kelas
  - Pengalokasian pengamatan ke dalam salah satu kelas
  - Pembuatan kotak (persegipanjang) pada setiap kelas dengan tinggi kotak masing-masing merupakan frekuensi

# HISTOGRAM

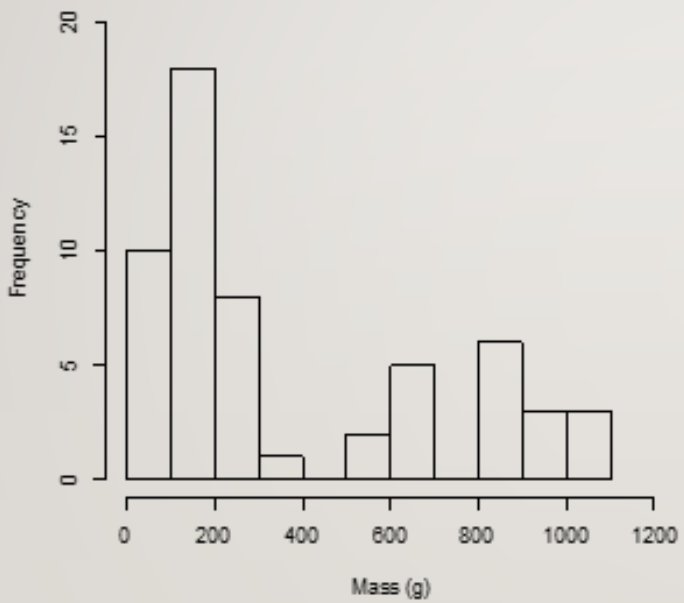
Data:

```
mass<-
c(5.9,32.0,40.0,51.5,70.0,100.0,78.0,80.0,85.0,85.0,110.0,115.0,125.0,130.0,1
20.0,120.0,130.0,135.0,110.0,130.0,150.0,145.0,150.0,170.0,225.0,145.0,188.0,
180.0,197.0,218.0,300.0,260.0,265.0,250.0,250.0,300.0,320.0,514.0,556.0,
840.0,685.0,700.0,700.0,690.0,900.0,650.0,820.0,850.0,900.0,1015.0,820.0,
1100.0,1000.0,1100.0,1000.0,1000.0)
```

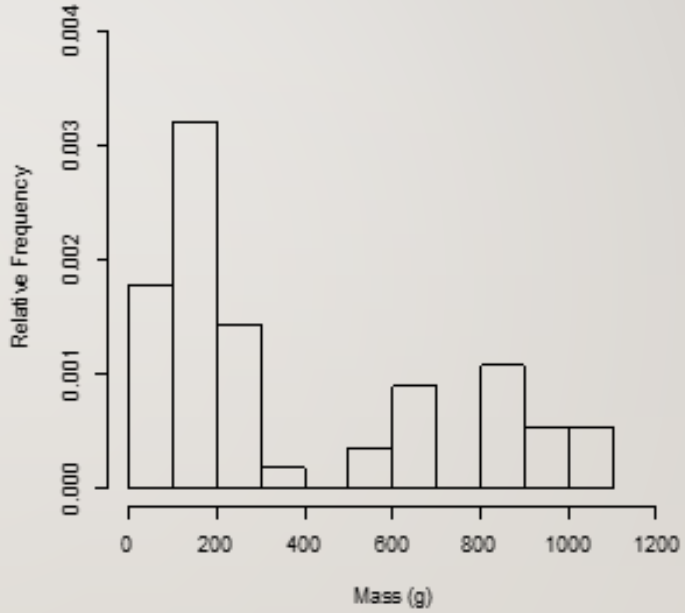
Program R untuk membentuk histogram frekuensi:

- `hist(mass,`**`freq=TRUE`**`,xlim=c(0,1200),breaks=12,`**`ylim=c(0,20),`** `main=NULL,xlab="Mass (g)")`
- `hist(mass,`**`freq=FALSE`**`,xlim=c(0,1200),breaks=12,`**`ylim=c(0,0.004),`** `main=NULL,xlab="Mass(g)",`**`ylab="Relative Frequency"`**`)`

# HISTOGRAM




hist(mass,freq=TRUE,xlim=c(0,1200),breaks=12,ylim=c(0,20),main=NULL,xlab="Mass (g)")




hist(mass,freq=FALSE,xlim=c(0,1200),breaks=12,ylim=c(0,0.004), main=NULL,xlab="Mass(g)",ylab="Relative Frequency")

# HISTOGRAM

- Beberapa aturan penentuan jumlah kelas:
  a) *Rule of twelve* (*ad hoc*): jumlah kelas ditetapkam sebanyak 12 yang merentang wilayah data.
  b) *Robust rule of twelve*: jumlah kelas ditetapkam sebanyak 12 yang merentang 4.45 IQR (*inter quartile range*) dari data atau merentang 6 simpangan baku (6σ) untuk sebaran normal yaitu sebanyak 99.7% di bawah kurva normal.
  c) *Sturge*: jumlah kelas ditetapkan berdasarkan populasi normal dengan formula berikut,

$$N_{ks} = 1 + \frac{\log(n)}{\log(2)}$$

# HISTOGRAM

- Beberapa aturan penentuan jumlah kelas:
  - *d) Doane*: aturan ini merupakan modifikasi dari aturan Sturge di mana jumlah kelas ditetapkan berdasarkan berdasarkan sebaran yang tidak simetrik dengan formula berikut,

$$N_{kd} = 1 + \frac{\log(n) + \log(1 + c_1)}{\log(2)}$$

$$c_1 = \frac{\mathrm{m}_3}{m_2^{3/2}} \left[ \frac{(n+1)(n+3)}{6(n-2)} \right]^{1/2}$$

$$m_j = \sum_{i=1}^{n} (X_i - \bar{X})^j \, / n$$

dengan j=2,3.

# HISTOGRAM

- Beberapa aturan penentuan jumlah kelas:
  e) *Scott*: jumlah kelas ditentukan dengan asumsi distribusi normal yang meminimumkan IMSE (*integrated mean square error*) antara histogram frekuensi relatif dan kepekatan peluang. Lebar kelasnya ($w_c$) adalah

$$w_c = 3.49sn^{-1/3}$$

  dengan s adalah simpangan baku dari data.

  f) *Freedman-Diaconis*: jumlah kelas ditentukan dengan meminimumkan MISE (*mean of the integrated square error*) antara histogram frekuensi relatif dan kepekatan peluang. Aturan ini menggunakan IQR sehingga lebih *robust* daripada aturan Scott. Lebar kelasnya ($w_c$) adalah

$$w_{fd} = 2\,IQR\,n^{-1/3}$$

  dengan IQR adalah wilayah antar kuartil dari data.

# HISTOGRAM

- Definisi Kelas:

  penentuan jumlah kelas dan batas kelas

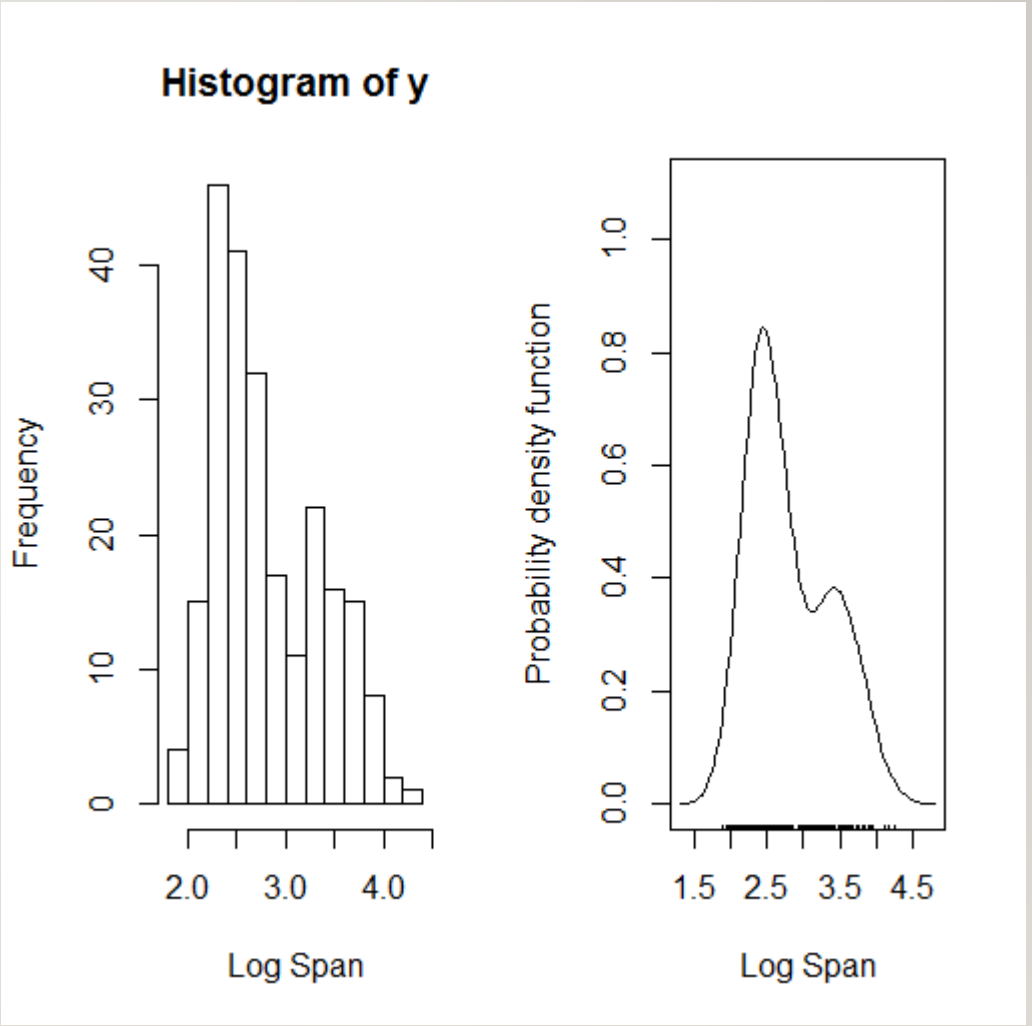
- Tidak Mulus (Not Smooth)

  fungsi *stepwise* meskipun fungsi kepekatan kontinu

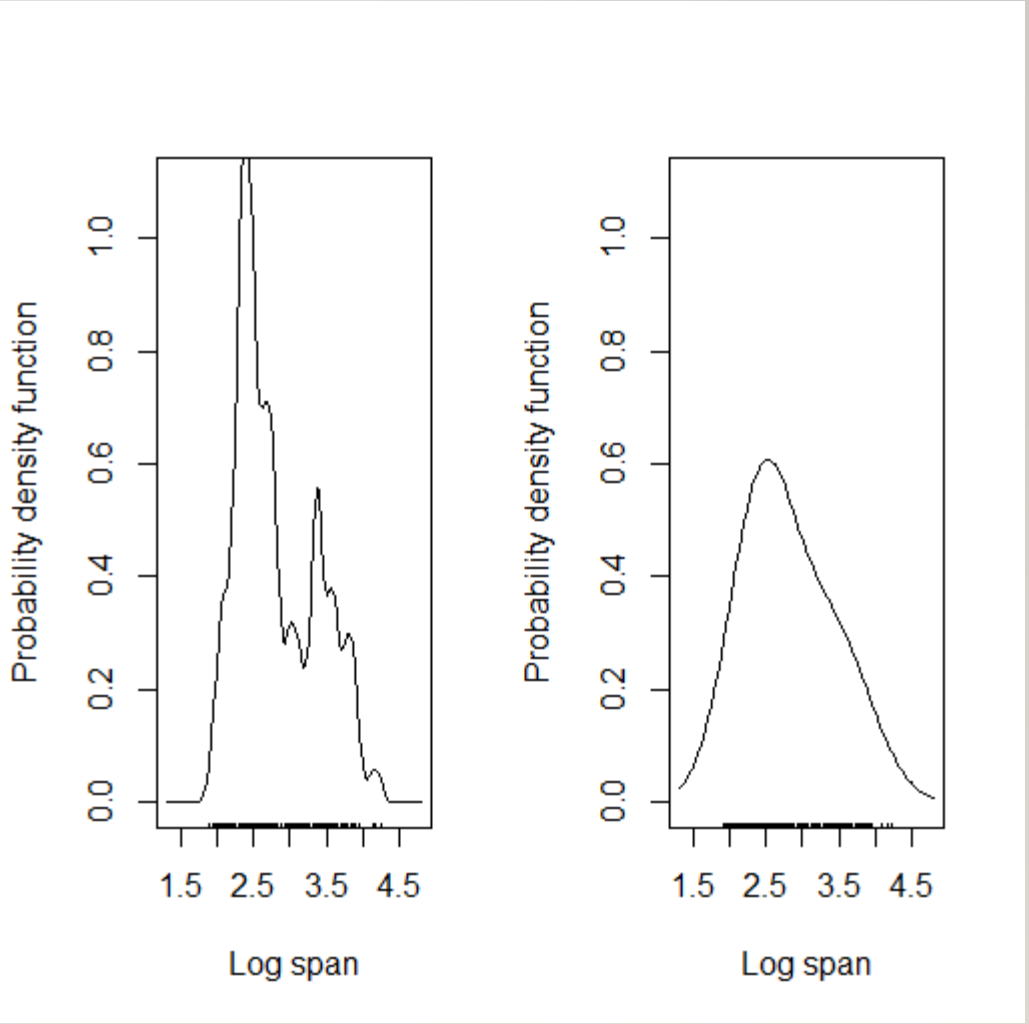
Solusi:

- Histogram lokal menghindari penentuan kelas
- Penduga kepekatan kernel agar kepekatannya mulus

```
library(sm)
y <- log(aircraft$Span[aircraft$Period==3])
par(mfrow=c(1,2))
hist(y, xlab="Log Span", ylab="Frequency")
sm.density(y, xlab="Log Span")
par(mfrow=c(1,1))
```



```
y <- log(aircraft$Span[aircraft$Period==3])
par(mfrow=c(1, 2))
sm.density(y, hmult = 1/3, xlab="Log span")
sm.density(y, hmult = 2, xlab="Log span")
par(mfrow=c(1,1))
```



```
y1 <- log(aircraft$Span[aircraft$Period==1])
y2 <- log(aircraft$Span[aircraft$Period==2])
y3 <- log(aircraft$Span[aircraft$Period==3])
sm.density(y3, xlab="Log span")
sm.density(y2, add=T, lty=2)
sm.density(y1, add=T, lty=3)
legend(3.5, 1, c("Period 1", "Period 2", "Period 3"), lty=1:3)
```