1.0 - Statistical modelling

Resources for this section

- Cf. "R basics" appendix for a brief recap of R basics
- So many resources available online... you could start with:
  - CRAN documents (manuals, contributed docs)
  - Hadley Wickham's online website
  - Garrett Grolemund (RStudio)'s Twitter account
- Go through the warm-up workout

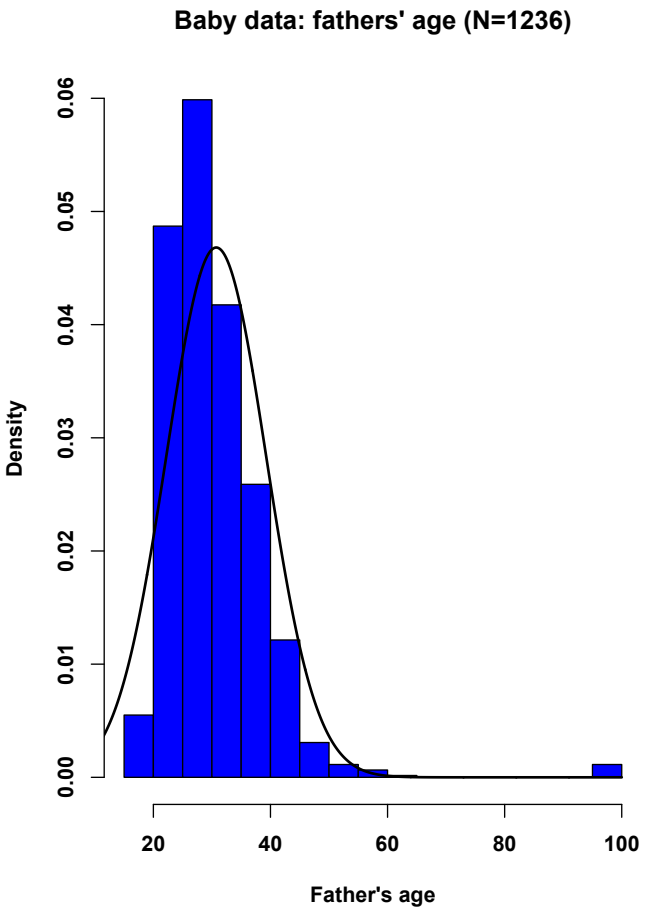
Motivation for modelling

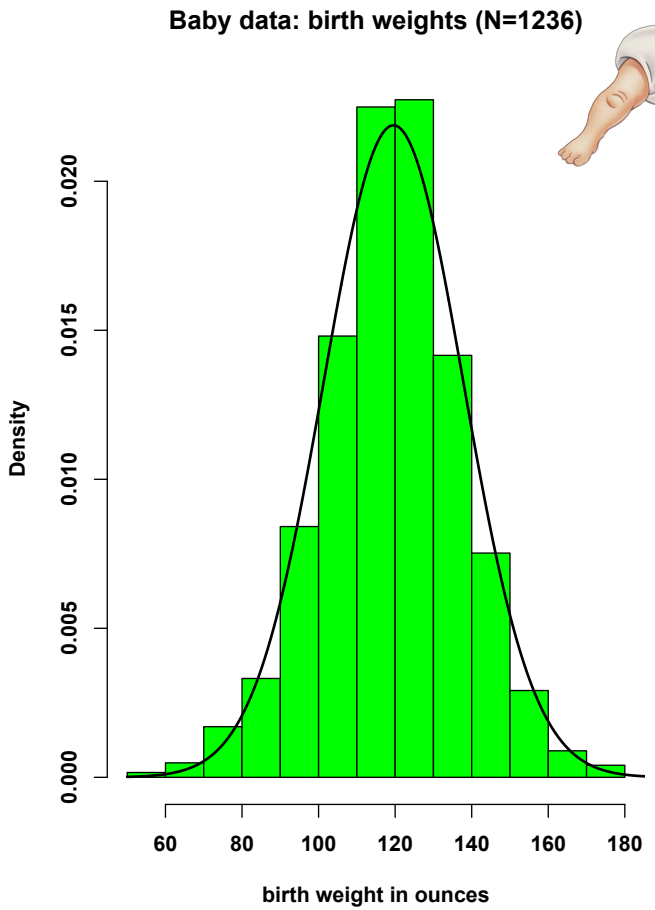
There are overall two kinds of objectives in modelling:

- Describe and characterize a random behaviour within the data
- Describe and characterize patterns of interest within the data

1.1 - Distribution models

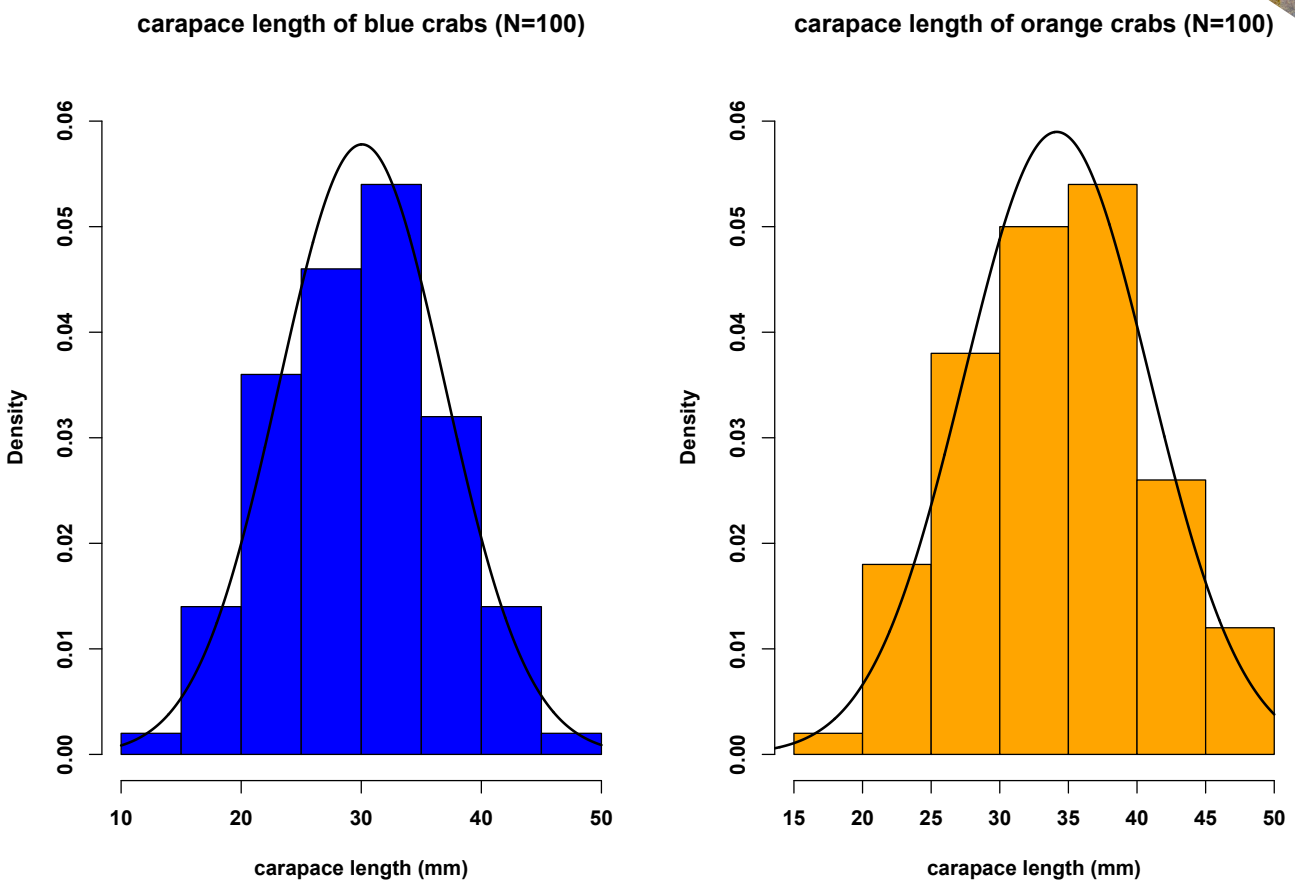
Modelling a random variable...





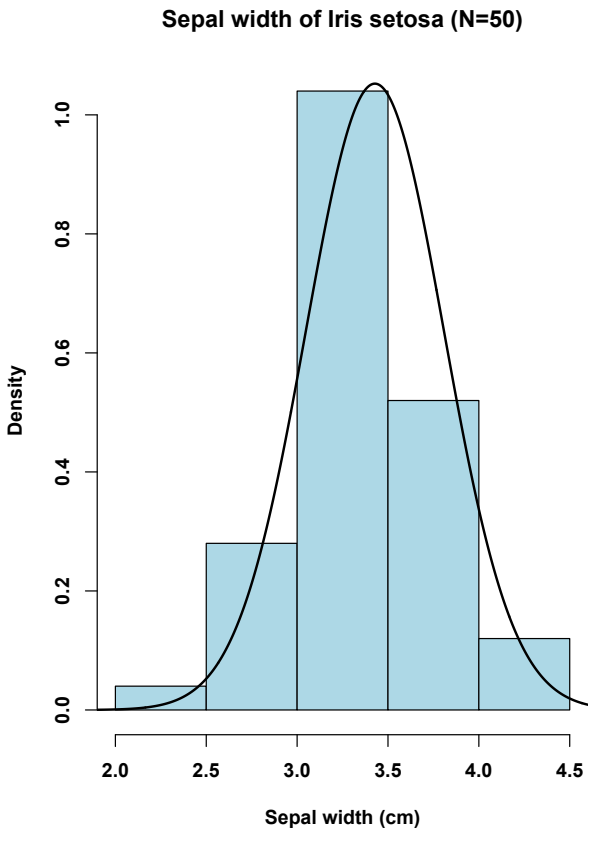
Dataset: http://www.stat.berkeley.edu/users/statlabs/labs.html.
Stat_Labs: Mathematical_Statistics_through_Applications_Springer-Verlag_(2001)_by_Deborah_Nolan_and_Terry_Speed.

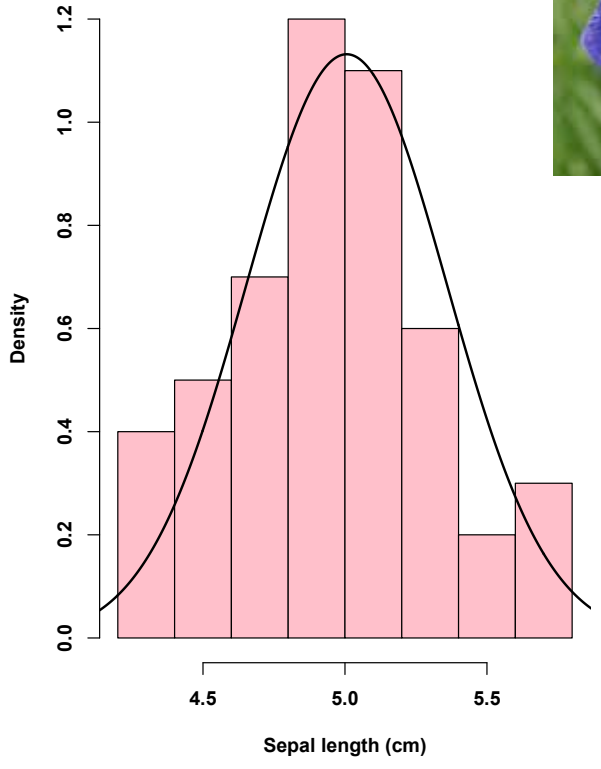
Modelling a random variable...

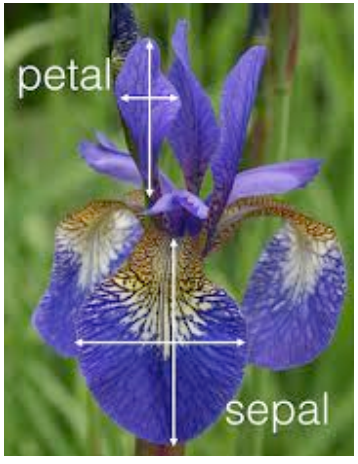


Campbell, N.A. and Mahon, R.J. (1974) A multivariate study of variation in two species of rock crab of genus Leptograpsus. Australian Journal of Zoology **22**, *417–425*

Modelling a random variable...







Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole.

Distributions in R

- Synopsis (where `distr` abbreviates a chosen distribution):
  `rdistr(n, ...)` - random generation of n realizations
  `ddistr(x, ...)` - probability density function $f(x)$
  `pdistr(q, ...)` - probability distribution function $F(x)$
  `qdistr(p, ...)` - quantile function
- Example with the Normal distribution:
  `rnorm(10)` randomly generates 10 realizations of $\mathcal{N}(0, 1)$
  `dnorm(0) == 1/ sqrt(2*pi)` should return `TRUE`
  `pnorm(1.645)` returns `0.9500151`
  `qnorm(0.95)` returns `1.644854`
- `?Distributions` for a list of available distributions: `dbeta`, `dbinom`, `dexp`, `dpois`, `dlnorm`, `dunif`, `dt`, `dcauchy`, etc.
- **See also:**
  http://cran.r-project.org/web/views/Distributions.html

Random generation

- `rnorm(2,1,3)` generates 2 realizations of $\mathcal{N}(\mu = 1, \sigma = 3)$
- The following loop:

  ```
  par(mfrow=c(1,3))
  for(i in 1:3){v=rnorm(50); hist(v)}
  ```

  plots the histograms of 3 samples of 50 realizations of $\mathcal{N}(0, 1)$
- This is how new samples can be generated in a loop
- `set.seed(s)` may be used to fix pseudo-random generation:

  ```
  set.seed(1); v=rnorm(10);
  set.seed(1); w=rnorm(10);
  ```

  (using seed `s=1`) will ensure that `v=w`
- This is particularly useful to **reproduce** simulated data exactly

Examining the distribution of a dataset

- `summary(object)` returns statistics for the input dataset:

  R> summary(rnorm(10))

  may return something like

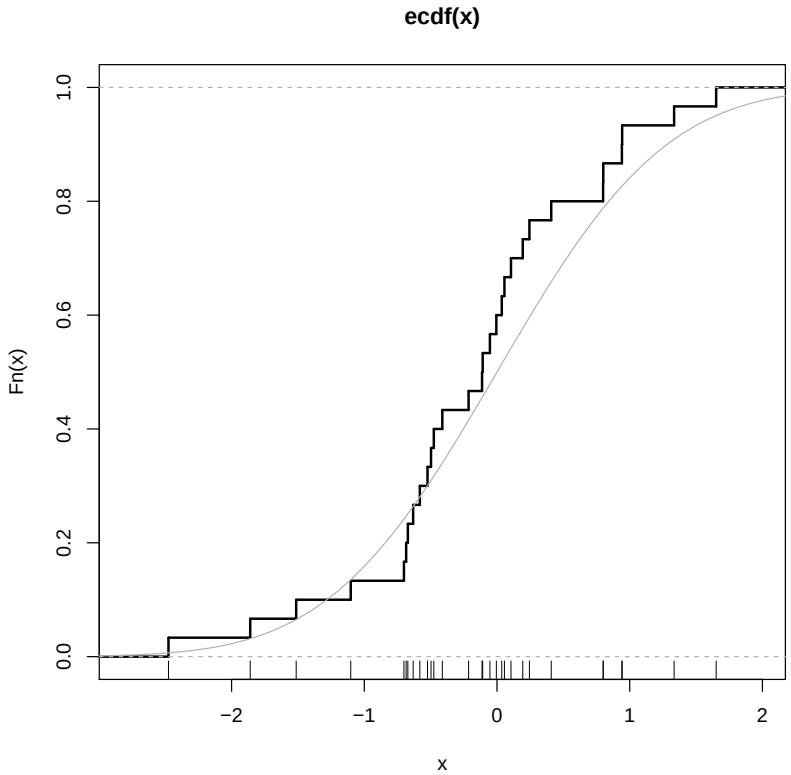
```
    Min.  1st Qu.  Median    Mean    3rd Qu.  Max.
-1.56300 -0.30350 -0.11910 -0.01992 0.44800 1.10500
```

- `summary()` also returns a summary of the output of many fitting functions
- Plots: `hist()`, `stem()`, `density()`, `boxplot()`...
- QQ-plots: `qqnorm()`, `qqplot()`
- `shapiro.test()` (Shapiro-Wilk normality test)

A simple example: empirical CDF

[from addictedtor.free.fr]

```
x <- rnorm(30)
F10 <- ecdf(x)
plot(F10, verticals= TRUE, do.p = FALSE, lwd=2)
curve(pnorm, from=-5, to=5, add=TRUE, col="gray70")
rug(x)      # plots the locations of x's below the curve
```



1.2 - Modelling associations

Statistical learning: regression

- Variable of interest: **Wage** (continuous)

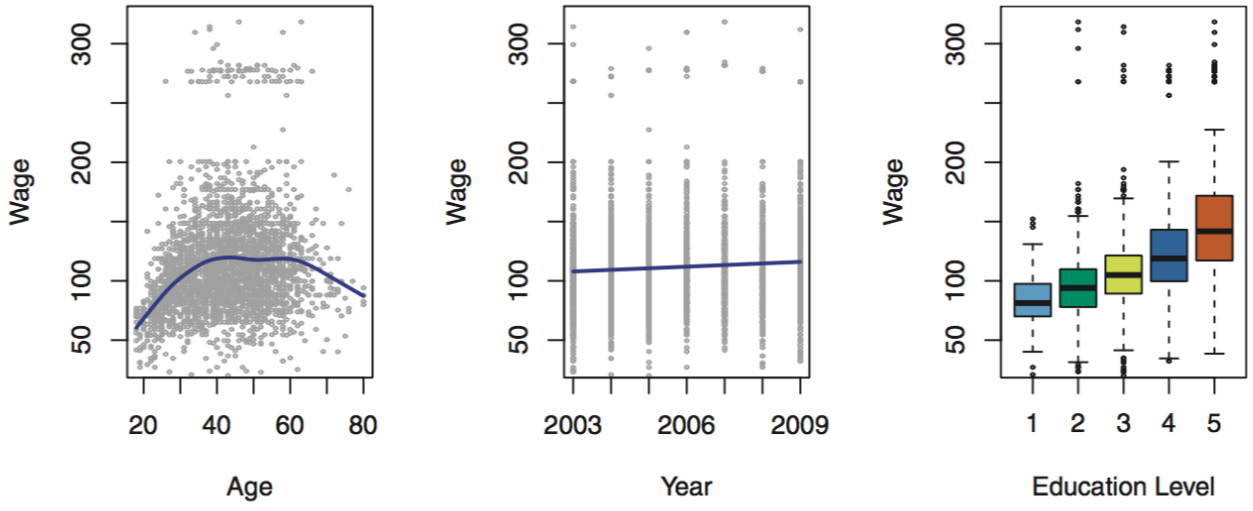


FIGURE 1.1. `Wage` *data, which contains income survey information for males from the central Atlantic region of the United States.* Left: `wage` *as a function of* `age`*. On average,* `wage` *increases with* `age` *until about* 60 *years of age, at which point it begins to decline.* Center: `wage` *as a function of* `year`*. There is a slow but steady increase of approximately* $10,000 *in the average* `wage` *between* 2003 *and* 2009*.* Right: *Boxplots displaying* `wage` *as a function of* `education`*, with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* `wage` *increases with the level of education.*

(figure from ISLR)

Statistical learning: classification

- Variable of interest: **Default** (categorical)

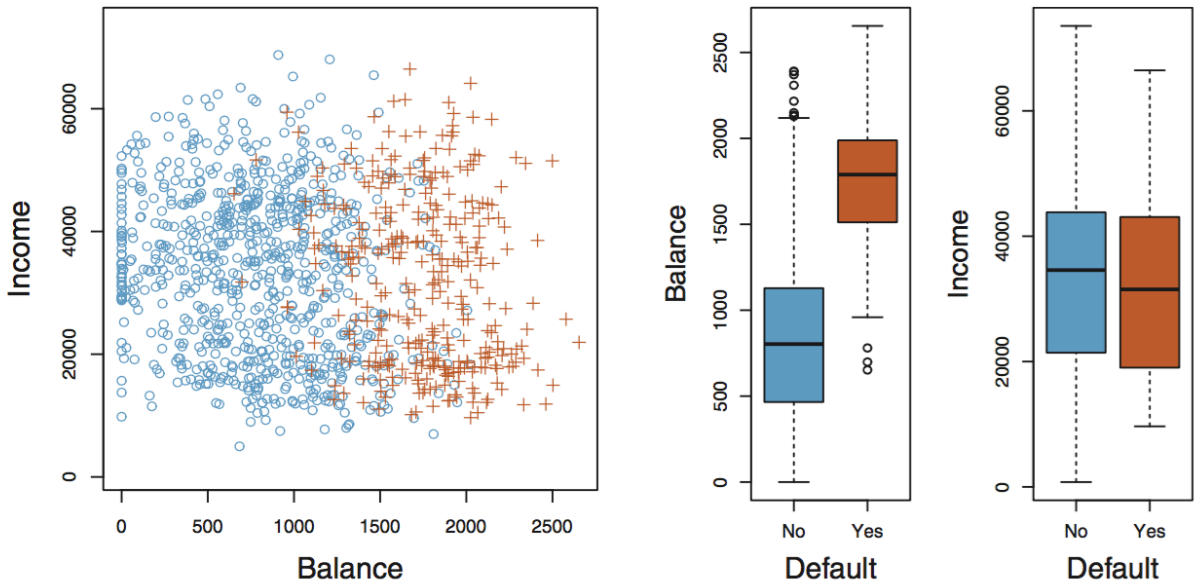
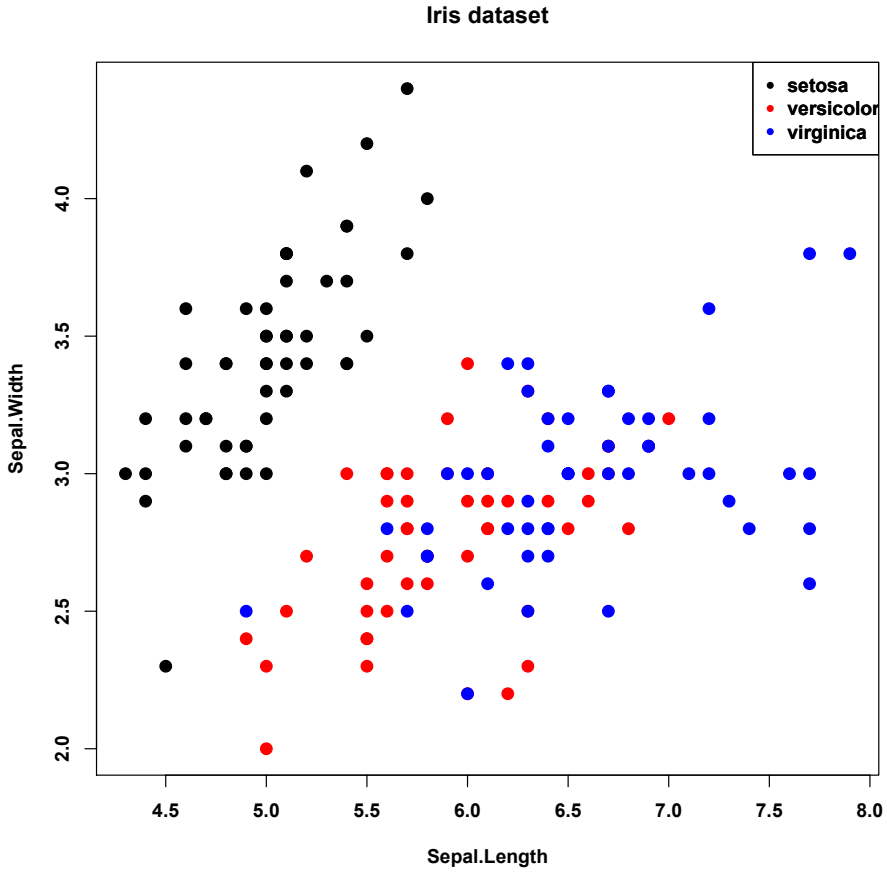


FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of* `default` *status.* Right: *Boxplots of* `income` *as a function of* `default` *status.*

(figure from ISLR)

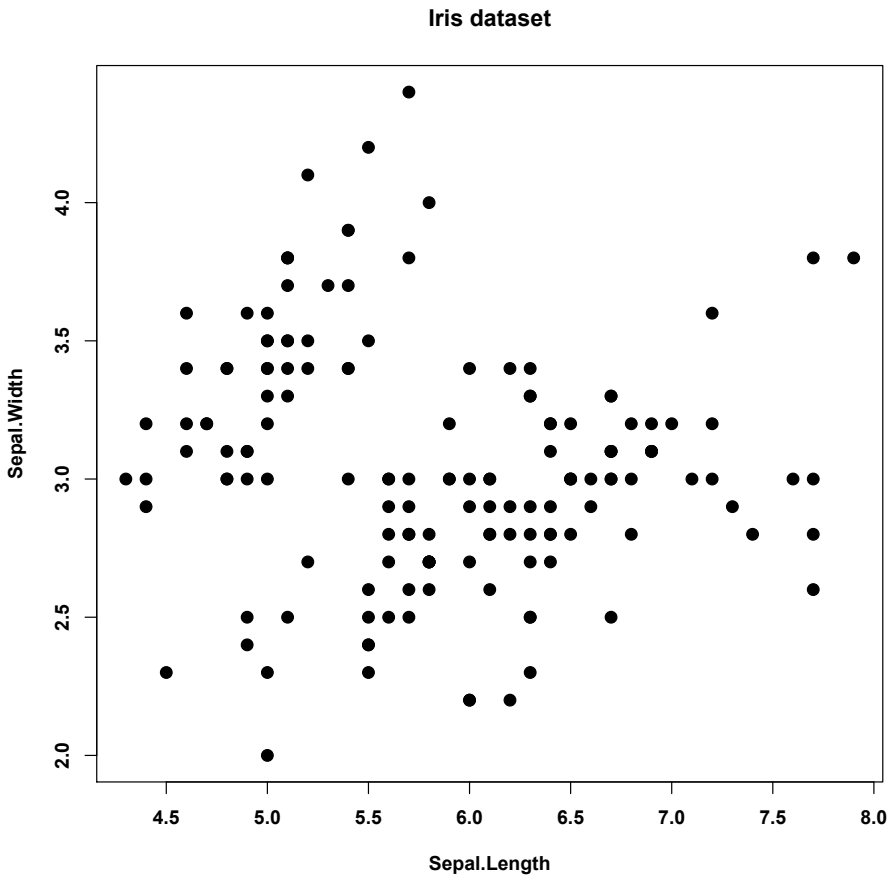
Statistical learning: supervised learning

- Variable of interest: **Species** (categorical)
- With prior experience available (**supervised classification**)



Statistical learning: unsupervised learning

- Variable of interest: **Species** (categorical)
- With no prior experience available (**clustering**)



1.3 - Linear regression

Linear regression is a central tool of statistical analysis. It is used extensively and often is a key component within a more complex analytical procedure. The approach was notably extended to generalized linear modelling (GLM) and nonlinear regression, to address a large range of problems.

Regression analysis is critical to a large number of methodological aspects, including:

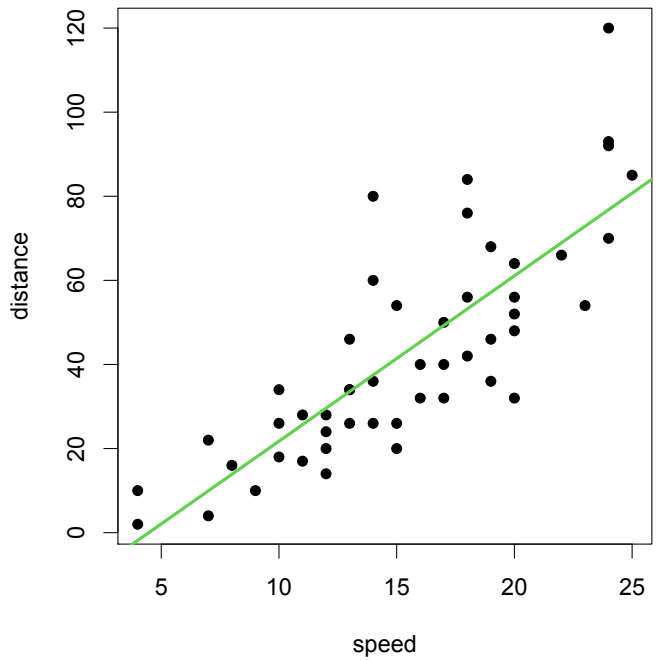
- trend estimation and prediction
- diagnostic tests for model selection
- benchmarking and (methodological) explorative analysis

Linear regression

In linear regression models, observations depend linearly on the parameters of interest:

$$Y = \theta_0 + \theta_1 X + \varepsilon$$

where ε are typically i.i.d. realizations of a Gaussian r.v.



Linear regression

In linear regression models, observations depend linearly on the parameters of interest:

$$Y = \theta_0 + \theta_1 X + \varepsilon$$

where ε are typically i.i.d. realizations of a Gaussian r.v.

- if $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, then $Y_i | X_i \sim \mathcal{N}(\theta_0 + \theta_1 X_i, \sigma^2)$
- Least squares are usually used to estimate $\theta = (\theta_0, \theta_1)$
- When the error term ε is not assumed Gaussian, maximum likelihood estimation is the more generic framework
- MLE = OLS when $\varepsilon \sim N(0, \sigma^2)$

Regression analysis with R

Linear regression via least squares is performed in R using `lm`:

- `lm(y~x)` implements fitting of $y = \beta_0 + \beta_1 x + \varepsilon$
- `lm(y~x+0)` implements fitting of $y = \beta_1 x + \varepsilon$

- Example: compare

```
summary(lm(dist~speed, data=cars))
summary(lm(dist~speed+0, data=cars))
```

- Example: the following Least Squares estimation

```
x = c(1:10); y = 3*x + rnorm(10); lm(y~x)
```

returns estimates $\hat{\beta}_0 = -0.8818$ and $\hat{\beta}_1 = 3.0975$

One can capture the output of `lm`, inspect it, and use it:

- `out = lm(y~x)  # capture output`
- `summary(out)  # inspect output`
- `plot(out$residuals)  # plot residuals from model fit`
- `acf(out$residuals)  # check these residuals!`
- etc.

1.4 - Extensions of the linear model

Generalised Linear Models (GLM)

- GLM is an extension of the simple linear model allowing for:
  - discrete response variable Y (e.g. Poisson observations)
  - non-Gaussian variable Y (by adapting the distribution used in the linear model via a link function)
  - categorical covariates X (e.g. binary variables)
- Example: $Y = \theta_1 X_1 + \theta_2 X_2 + \varepsilon$, where
  - $Y \in \mathbb{R}$ patient survival time
  - $X1 =$ white blood cell counts
  - $X2 =$ presence of in leukemia patients
- In R:

```
library(survival) # contains the leuk dataset
summary(glm(time~log(wbc)+ag, family="Gamma",
                    data=leuk))
```

Logistic regression

- Logistic regression models binary response variables Y
- Particular case of GLM: it allows for categorical covariates
- Logistic regression is therefore a classification model

- Example: $Y = \theta_1 X_1 + \theta_2 X_2 + \varepsilon$, where
  - $Y \in \{0; 1\}$, $Y = 0$ if subject is a blue crab, $Y = 1$ if orange
  - $X_1 =$ carapace length (CL, in mm)
  - $X_2 =$ carapace width (CW, in mm)
- In R (`sp` corresponds to crab species Y):

```
library(MASS) # contains the crabs dataset
summary(glm(sp~CL+CW, family="binomial", data=crabs))
```

1.5 - Nonparametric density estimation

Nonparametric Statistics?

Nonparametric may mean that:

- the method does not assume an underlying distribution;
  - Ex: rank statistics
- the model is allowed to change with the data.
  - Ex: nonparametric regression

Some nonparametric methods are also sometimes termed *adaptive*, because they adapt to the underlying distribution of the data.

Nonparametric density estimation

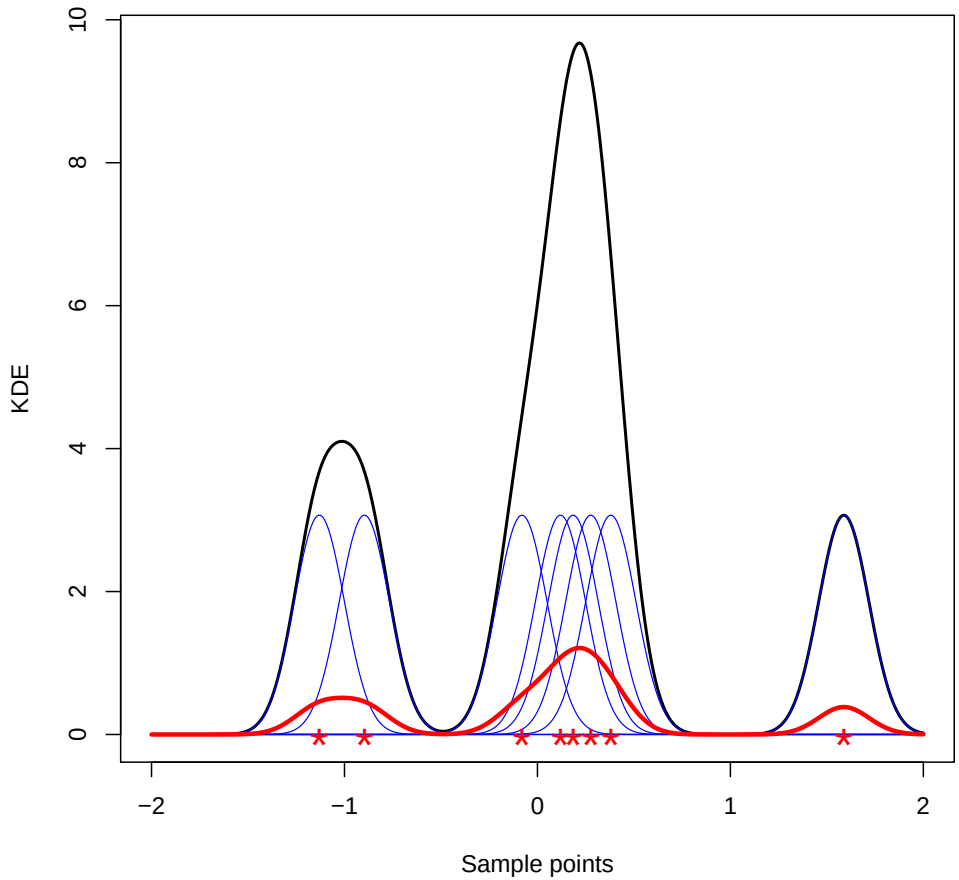
- Estimate a density $f(x)$ with a smooth object $\hat{f}(x)$
- Use e.g. a Kernel Density Estimator (KDE)
- Consider sample points as "kernels"
- Centre a given "template" density of size h at each kernel:

$$K_h(u - x_i) = \frac{1}{h} K\left(\frac{u - x_i}{h}\right)$$

- Sum contributions of all these kernel densities:

$$\hat{f}(u) = \frac{1}{N} \sum_{i=1}^{N} K_h(u - x_i)$$

Kernel Density Estimation



$x_i(*), \quad K\left(\frac{x-x_i}{h}\right), \quad \frac{1}{h}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right), \quad \frac{1}{N}\sum_{i=1}^{N} K_h(x-x_i)$

(example with $N = 8$ points)

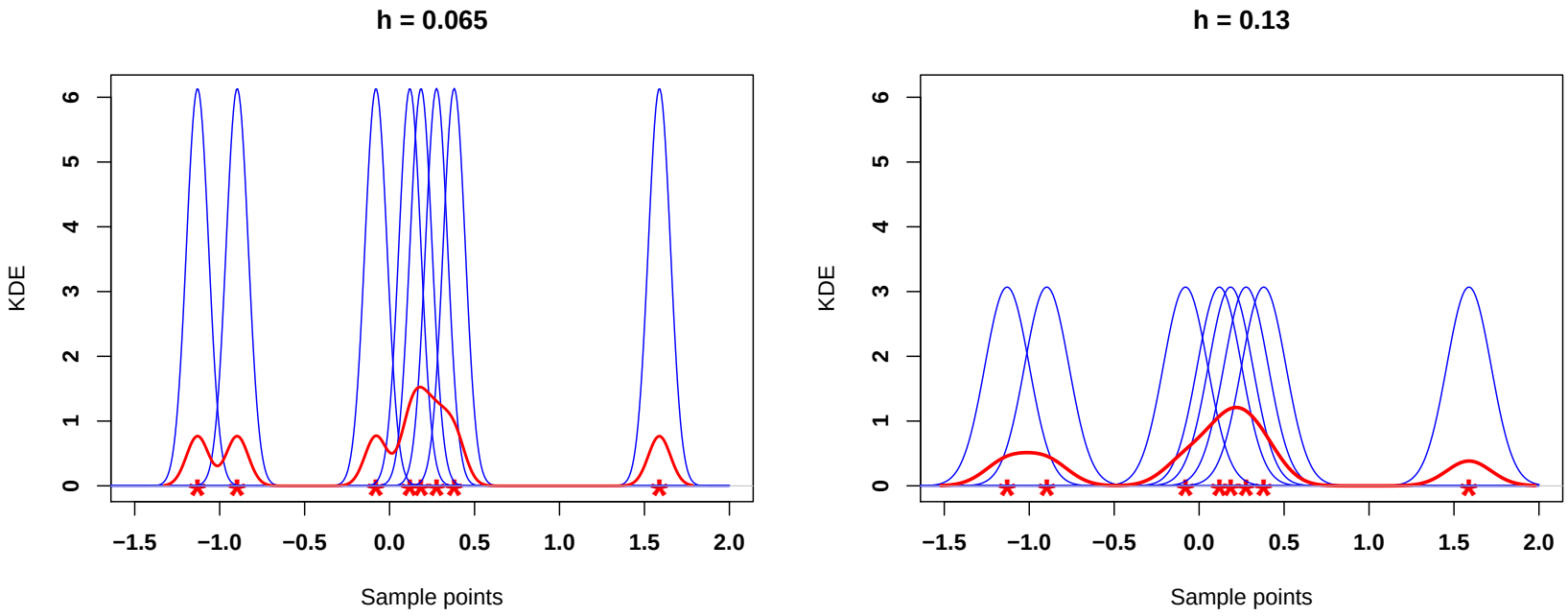
Kernel Density Estimation

- In R, estimate the pdf of a sample `x` by `density(x)`
- Example:

```
xs = rnorm(10)
fx = density(xs)
names(fx)
plot(fx)
points( density(xs, bw=2*fx$bw) )
lines(fx$x, fx$y, col='blue', lwd=2)
```

Kernel Density Estimation

Visualize influence of choice of bandwidth h:
kernel densities are in blue, final density estimate is in red...



(example with $N = 8$ points)

Other common nonparametric density estimators

- Naive estimators using $F_N(u) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{X_i \leq u}$:

$$f_N(u) = \frac{1}{Nh}\sum_{i=1}^{N} W\left(\frac{u - X_i}{h}\right)$$

  where $W(u) = \frac{\delta(|u|<1)}{2}$ is a rectangular kernel

- k-th nearest neighbour estimators ($k \geq 2$)

$$\hat{f}_N(u) = \frac{k}{2Nd_k(u)} = \frac{1}{Nd_k(u)}\sum_{i=1}^{N} K\left(\frac{u - X_i}{d_k(u)}\right)$$

  where $d_k(u) = |u - X_{[k]}|$ are in increasing order, and $\int K(u)du = 1$