

Gaussian distribution

Maura Mezzetti

`maura.mezzetti@uniroma2.it`

Sufficient Statistics

$$f(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

$$f(y_1, \dots, y_i, \dots, y_n|\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

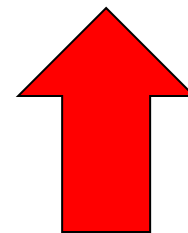
$$f(y_1, \dots, y_i, \dots, y_n|\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2}$$

σ^2 known, Sufficient Statistics for μ

$$f(y_1, \dots, y_i, \dots, y_n \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2}$$

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

$$f(y \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2} e^{+\frac{1}{\sigma^2} \mu \sum_{i=1}^n y_i} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n \mu^2}$$



σ^2 known, Sufficient Statistics for μ

$$f(y_1, \dots, y_i, \dots, y_n \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2}$$

$$f(y \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2} e^{+\frac{1}{\sigma^2} \mu \sum_{i=1}^n y_i} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n \mu^2}$$

$$T(y) = \sum_{i=1}^n y_i \text{ sufficient statistics for } \mu$$

Sufficient Statistics for σ^2

$$f(y_1, \dots, y_i, \dots, y_n \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2}$$

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

$$f(y \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2} e^{+\frac{1}{\sigma^2} \mu \sum_{i=1}^n y_i} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n \mu^2}$$

Sufficient Statistics for σ^2

$$f(y_1, \dots, y_i, \dots, y_n \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2}$$

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

$$f(y \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2} e^{+\frac{1}{\sigma^2} \mu \sum_{i=1}^n y_i} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n \mu^2}$$

Sufficient Statistics for σ^2

$$f(y_1, \dots, y_i, \dots, y_n \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2}$$

$$f(y \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2} e^{+\frac{1}{\sigma^2} \mu \sum_{i=1}^n y_i} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n \mu^2}$$

$$T(y) = \left(\sum_{i=1}^n y_i, \sum_{i=1}^n y_i^2 \right) \text{ sufficient statistics for } \sigma^2$$

Suppose that we have two variables

1. Y – the dependent variable (response variable)
2. X – the independent variable (explanatory variable, factor)-

X may or may not be a random variable .

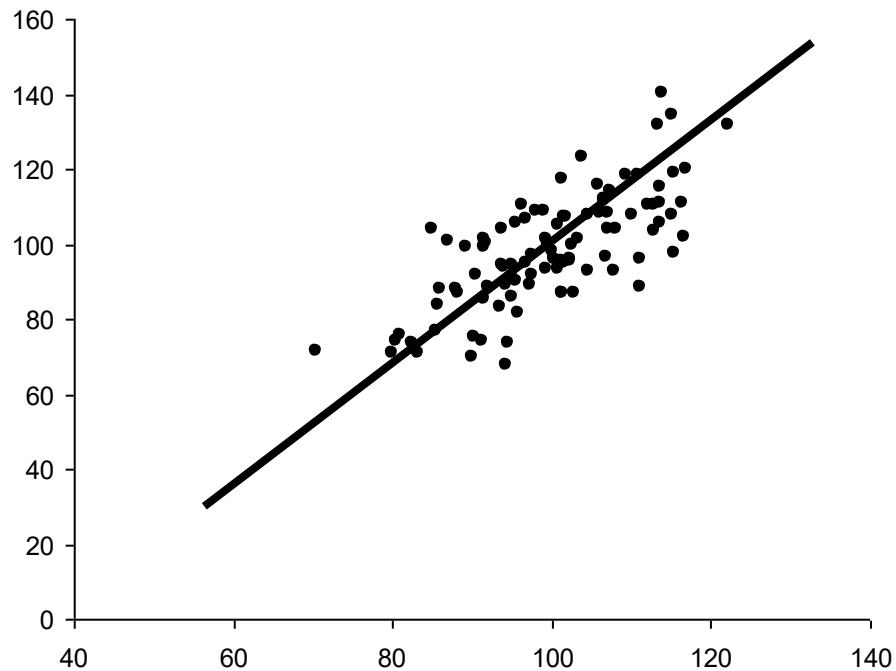
The dependent variable, Y , is assumed to be a random variable. The distribution of Y is dependent on X

Assume that we have collected data on two variables X and Y . Let

$$(X_1, Y_1) (X_2, Y_2) (X_3, Y_3) \cdots (X_n, Y_n)$$

denote the pairs of measurements on the on two variables X and Y for n cases in a sample (or population)

- When data is correlated it falls roughly about a straight line.



Regression Line

At a given value of x , the equation:

$$\hat{y} = \beta_0 + \beta_1 x$$

Predicts a single value of the response variable

But... we should not expect all subjects at that value of x to have the same value of y

Variability occurs in the y values!

Simple Linear Regression Model

- $y = \beta_0 + \beta_1 x + \varepsilon$
 - x : regressor variable
 - y : response variable
 - β_0 : the intercept, unknown
 - β_1 : the slope, unknown
 - ε : error with $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2$ (unknown)
- The errors are uncorrelated.

Linear Regression Model

- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
 - x_i : regressor variable
 - y_i : response variable
 - β_0 : the intercept, unknown
 - β_1 : the slope, unknown
 - $\varepsilon_i \sim N(0, \sigma^2)$
- The errors are uncorrelated.

Sufficient Statistics

$$\begin{aligned} & f(y_1, \dots, y_i, \dots, y_n | x_1, \dots, x_n, \mu, \sigma^2) \\ &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - [\beta_0 + \beta_1 x_i])^2} \\ &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i^2 - 2y_i[\beta_0 + \beta_1 x_i] + [\beta_0 + \beta_1 x_i]^2)} \\ &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} (\sum_{i=1}^n y_i^2 - 2\beta_0 \sum_{i=1}^n y_i - 2\beta_1 \sum_{i=1}^n y_i x_i + n\beta_0^2 + 2\beta_0 \beta_1 \sum_{i=1}^n x_i + \beta_1^2 \sum_{i=1}^n x_i^2)} \end{aligned}$$

Sufficient Statistics

$$\begin{aligned}
 & f(y_1, \dots, y_i, \dots, y_n | x_1, \dots, x_n, \mu, \sigma^2) \\
 &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - [\beta_0 + \beta_1 x_i])^2} \\
 &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i^2 - 2y_i[\beta_0 + \beta_1 x_i] + [\beta_0 + \beta_1 x_i]^2)} \\
 &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{n\beta_0^2}{2\sigma^2}} e^{-\frac{1}{2\sigma^2} (\sum_{i=1}^n y_i^2 - 2\beta_0 \sum_{i=1}^n y_i - 2\beta_1 \sum_{i=1}^n y_i x_i + n\beta_0^2 + 2\beta_0 \beta_1 \sum_{i=1}^n x_i + \beta_1^2 \sum_{i=1}^n x_i^2)} \\
 &= h(x, y) c(\theta) \exp \left(\sum_{j=1}^k w_j(\theta) t_j(x, y) \right)
 \end{aligned}$$

$$T_1 = \sum_{i=1}^n y_i^2 \quad T_2 = \sum_{i=1}^n y_i \quad T_3 = \sum_{i=1}^n y_i x_i \quad T_4 = \sum_{i=1}^n x_i \quad T_5 = \sum_{i=1}^n x_i^2$$

Sufficient Statistics

$$f(y_1, \dots, y_i, \dots, y_n | x_1, \dots, x_n, \mu, \sigma^2) =$$

$$= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} (\sum_{i=1}^n y_i^2)} e^{-\frac{1}{2\sigma^2} (-2\beta_0 \sum_{i=1}^n y_i - 2\beta_1 \sum_{i=1}^n y_i x_i + n\beta_0^2 + 2\beta_0\beta_1 \sum_{i=1}^n x_i + \beta_1^2 \sum_{i=1}^n x_i^2)}$$

$$T_1 = \sum_{i=1}^n y_i \quad T_2 = \sum_{i=1}^n y_i x_i \quad T_3 = \sum_{i=1}^n x_i \quad T_4 = \sum_{i=1}^n x_i^2$$