Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

457.212 Statistics for Civil & Environmental Engineers
In-Class Material: Class 26 (end-of-semester)
Regression Analysis (2) (A&T: 8.4-8.7) & Extreme Value Distribution (A&T 4.2.3)

1. Multiple Linear Regression

"Linear regression of Y on $X_1, \ldots, X_m$"

(a) Define Δ^2 by assuming $\sigma^2_{Y|x} = \sigma^2$ (constant) or $\sigma^2_{Y|x} = \sigma^2 g^2(x_1,...,x_m)$ (non-constant)

(b) Find

$$\mathrm{E}[Y \mid x_1,...,x_m] = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m$$

(c) Estimate $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_m$ by solving

$$\frac{\partial \Delta^2}{\partial \beta_0} = \frac{\partial \Delta^2}{\partial \beta_1} = \cdots = \frac{\partial \Delta^2}{\partial \beta_m} = 0$$

(d) $s^2_{Y|x_1,...,x_m} = \dfrac{\Delta^2}{n-m-1}$

(**Note**: $m = 1$ for single linear regression)

2. Nonlinear Regression & Applications of Regression Analysis (Read A&T 8.6-8.7)

3. Correlation Analysis

(a) (True or theoretical) correlation coefficient

$$\rho_{XY} = \frac{Cov[X,Y]}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[XY] - \mu_X \mu_Y}{\sigma_X \sigma_Y}$$

(a) Unbiased estimator of ρ_{XY}, $\hat{\rho}$

$$\hat{\rho} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} x_i y_i - n\bar{X}\bar{Y}}{s_x s_y}$$

(b) $\hat{\rho}$ and $\hat{\beta}$

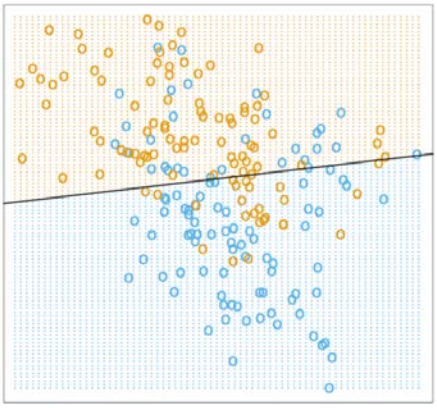
$$\hat{\rho} = \hat{\beta} \frac{s_X}{s_X} = \frac{\Sigma(x_i - \bar{X})(y_i - \bar{Y})}{(n-1)s_X^2} \frac{s_X}{s_Y} = \frac{\Sigma(x_i - \bar{X})(y_i - \bar{Y})}{\Sigma(x_i - \bar{X})^2} \frac{s_X}{s_Y} = \hat{\beta} \frac{s_X}{s_Y}$$

Instructor: Junho Song
junhosong@snu.ac.kr

(d) $\hat{\rho}^2$ and $r^2 = 1 - s_{Y|x}^2 / s_Y^2$

$$\hat{\rho}^2 = 1 - \frac{n-2}{n-1}\frac{s_{Y|x}^2}{s_Y^2}. \qquad \text{As } n \to \infty,\ \hat{\rho}^2 \to 1 - \frac{s_{Y|x}^2}{s_Y^2} = r^2$$

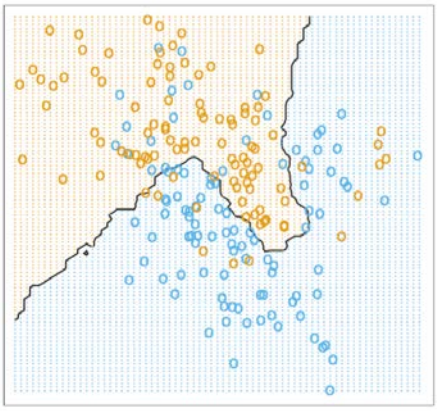
4. "Model-based" vs "Data-based" prediction



(a) Model-based prediction: assumes a smooth model and fit

e.g. Linear regression $Y = \beta_0 + \sum_{i=1}^{N} \beta_i x_i$

- maybe inaccurate, but stable



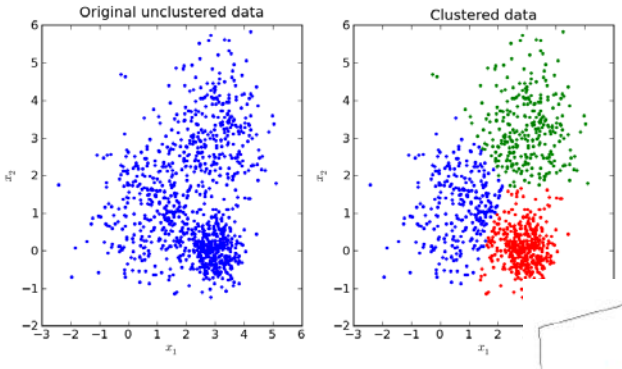
(b) Data-based prediction (interpolation): does not assume a model, just interpolate from adjacent data points

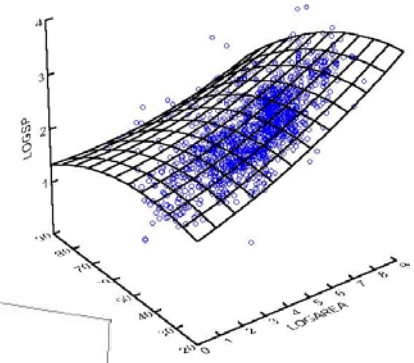
e.g. *k*-nearest neighbor model $Y = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$

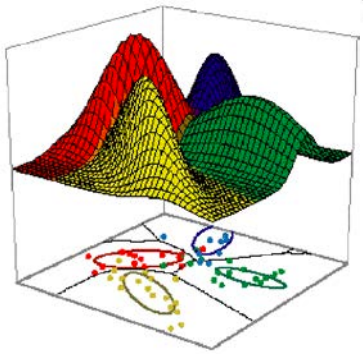
- Accurate, but may be unstable

5. Statistical/Machine Learning

Build a prediction model by (1) Clustering, (2) Classification, and (3) Regression







Instructor: Junho Song
junhosong@snu.ac.kr

6. Extreme Value Distributions

Given: The distribution model of a random quantity X, i.e. PDF or CDF

Question: From a sample of size n, the distribution of the minimum or maximum?

→ Deriving an ( ) distribution

e.g. maximum flood (or drought) in the next 100 years, maximum traffic load on bridge in the next 50 years

(More generally, the distribution of the *k*-th largest or smallest from a sample → "_____ Statistics")

(a) Deriving "**Exact Distributions**"

- Maximum: $Y_n = \max\left(X_1, X_2, ..., X_n\right)$

Under the assumption that $X_1, X_2, ..., X_n$ are statistically ( ) and ( ) distributed,

$$F_{Y_n}(y) = P(X_1 \le y \quad X_2 \le y \quad ... X_n \le y)$$
$$= [\quad\quad]^n$$

The corresponding PDF is therefore,

$$f_{Y_n}(y) = \frac{dF_{Y_n}(y)}{dy} = n[\quad\quad]^{n-1} f_X(y)$$

- Minimum: $Y_1 = \min\left(X_1, X_2, ..., X_n\right)$

$$1 - F_{Y_1}(y) = P(X_1 > y \quad X_2 > y \quad ... X_n > y)$$
$$= [\quad\quad]^n$$

Therefore, the CDF of Y_1 is

$$F_{Y_1}(y) = 1 - [\quad\quad]^n$$

The corresponding pdf is

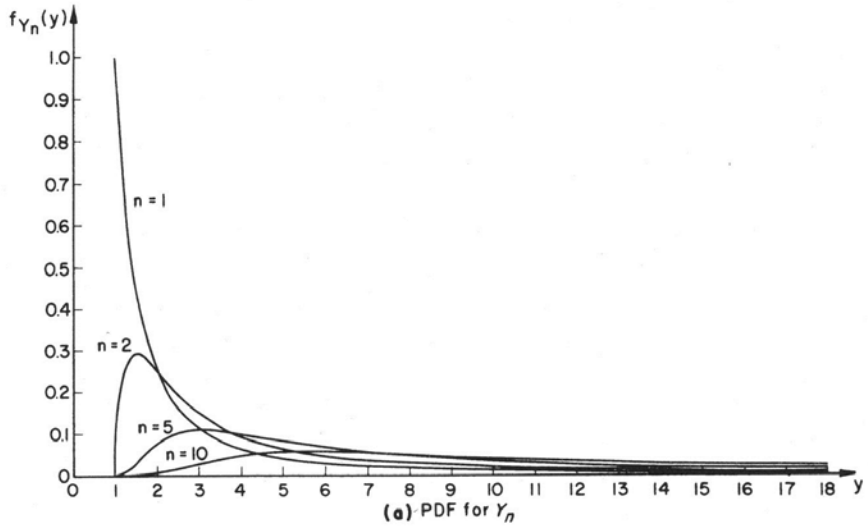
$$f_{Y_1}(y) = \frac{dF_{Y_1}(y)}{dy} = n[\quad\quad]^{n-1} f_X(y)$$

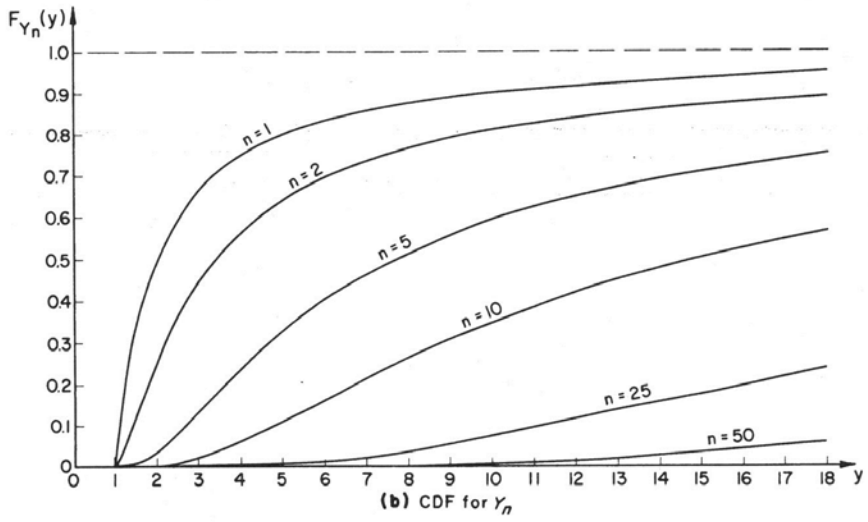
Instructor: Junho Song
junhosong@snu.ac.kr

Example 1: Suppose the PDF of a random variable X is given as below.

$$f_X(x) = \frac{1}{x^2}, \quad x \geq 1$$

When someone constructs a sample of size n, derive the CDF and PDF of the largest in the sample, i.e. $Y_n = \max\left(X_1, X_2, ..., X_n\right)$.

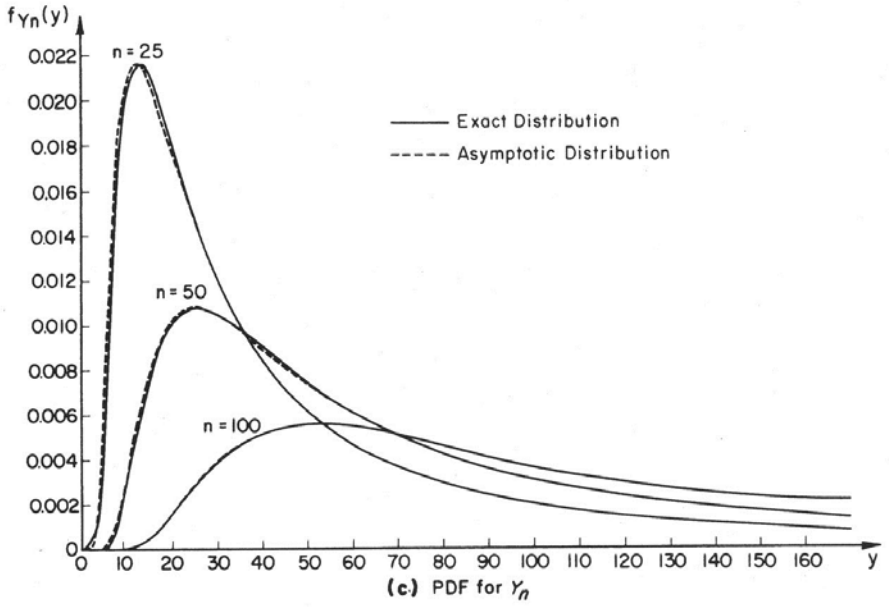




Instructor: Junho Song
junhosong@snu.ac.kr

(b) Asymptotic Distributions

An asymptotic distribution can be derived for large samples, i.e. $n \to \infty$, using Cramer's method (1946). For the example above, the exact (i.e. derived) and asymptotic distributions are compared as follows.



The asymptotic distributions of the extremes tend to converge on certain limiting forms (Gumbel 1958):

- **Type I**: The ( ) exponential form, $\exp\left[-e^{-A(n)y}\right]$
  - Gumbel distribution (largest)
- **Type II**: The exponential form, $\exp\left[-A(n)/y^{k}\right]$
  - Fisher-Tippett distribution (largest)
- **Type III**: The exponential form with upper/lower bound, $\exp\left[-A(n)/(\omega-y)^{k}\right]$
  - Weibull distribution (smallest)

The type is determined by the ( ) behavior of the original probability density function.

- Exponentially decaying tail (e.g. Normal) → **Type I**
- Polynomial tail (e.g. Example 3) → **Type II**
- Polynomial tail with the limited extreme value → **Type III**