Applications (사례 연구)

Chanseok Park (박찬석)

Applied Statistics Laboratory
Department of Industrial Engineering
Pusan National University

August 5, 2020

Hosted by SEC

부산대학교
PUSAN NATIONAL UNIVERSITY

Overview

1. Robust statistics와 응용 사례
   - Robust statistics (Basic applications)
   - Robust t-test
   - Robust design with contaminated data
2. Missing · Incomplete Data와 응용 사례
   - RD with unbalanced samples
   - RD with incomplete data
   - Competing risks with censoring, masking, etc.
   - Load-sharing
   - Grouped Data
3. Future work (missing with contamination)

Overview

1. Robust statistics와 응용 사례
   - Robust statistics (Basic applications)
   - Robust t-test
   - Robust design with contaminated data
2. Missing · Incomplete Data와 응용 사례
   - RD with unbalanced samples
   - RD with incomplete data
   - Competing risks with censoring, masking, etc.
   - Load-sharing
   - Grouped Data
3. Future work (missing with contamination)

Overview

1. Robust statistics와 응용 사례
   - Robust statistics (Basic applications)
   - Robust t-test
   - Robust design with contaminated data
2. Missing · Incomplete Data와 응용 사례
   - RD with unbalanced samples
   - RD with incomplete data
   - Competing risks with censoring, masking, etc.
   - Load-sharing
   - Grouped Data
3. Future work (missing with contamination)

Basic applications (estimating μ and σ)

Consider observations from $X_i \sim N(\mu, \sigma^2)$. We need to estimate μ and σ^2.

- MLE (maximum likelihood estimator)

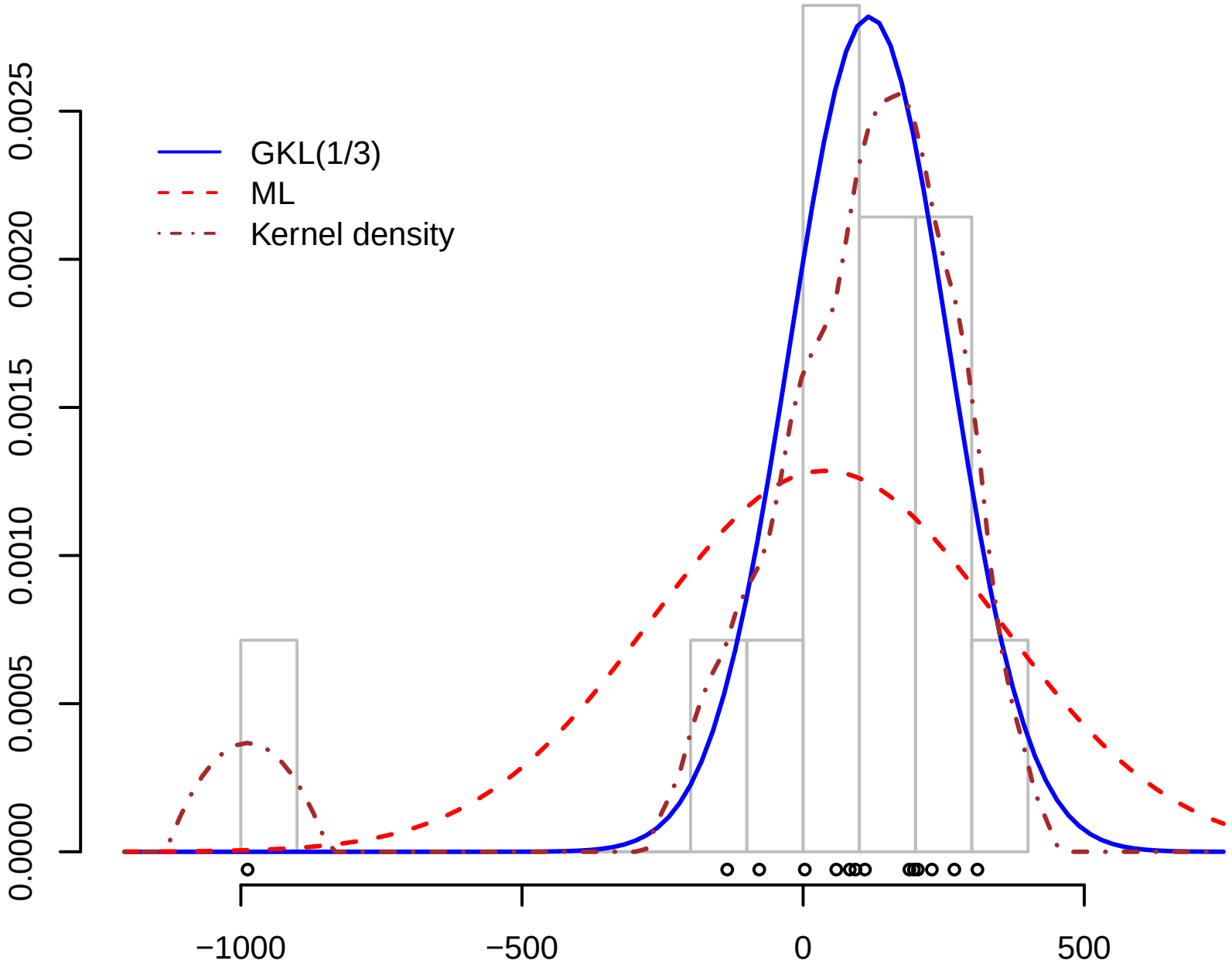
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

- BUE (best unbiased estimator) or UMVUE

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{\boldsymbol{n-1}}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- MDE (minimum distance estimator): KL, GKL, etc.
  MLE is a special case of GKL (Basu et al., 2011) which can have robustness. MDE is asymptotically fully efficient, but its calculation is quite complex. Thus, HL and Shamos are recommended (Talk-2).

1. Robust statistics와 응용 사례: Basic applications



1. Robust statistics와 응용 사례: Robust t-test

Darwin (1876) collected the data: the growth of pairs of corn (especially Zea May) seedings, one produced by self-fertilization and the other produced by cross-fertilization. For the data set, see Friendly et al. (2018).

Cross	23.500	12.000	21	22	19.125	21.500	22.125	20.375
	18.25	21.625	23.25	21	22.125	23.0	12	
Self	17.375	20.375	20	20	18.375	18.625	18.625	15.250
	16.50	18.000	16.25	18	12.750	15.5	**18**	
Difference	6.125	−8.375	1.000	2.000	0.750	2.875	3.500	5.125
	1.750	3.625	7.000	3.000	9.375	7.500	−6.000	

We can test $H_0 : \mu_x = \mu_y$ and $H_1 : \mu_x \neq \mu_y$, equivalently, $H_0 : \mu_d = 0$ and $H_1 : \mu_d \neq 0$, where $\mu_d = \mu_x - \mu_y$. A typical paired sample t-test with

$$T = \frac{\bar{D} - 0}{S_D/\sqrt{n}},$$

where $\bar{D}$ and S_D are the sample mean and standard deviation, which this becomes a one-sample t-test.

1. Robust statistics와 응용 사례: Robust t-test

Theorem 1 (Park, 2018a)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Then we have

$$T_A = \sqrt{\frac{2n}{\pi}}\Phi^{-1}\left(\frac{3}{4}\right)\frac{\underset{1\le i\le n}{\text{median}}\, X_i - \mu}{\underset{1\le i\le n}{\text{median}}\left|X_i - \underset{1\le i\le n}{\text{median}}\, X_i\right|} \xrightarrow{d} N(0,1). \tag{1}$$

Theorem 2 (Jeong et al., 2018)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Then we have

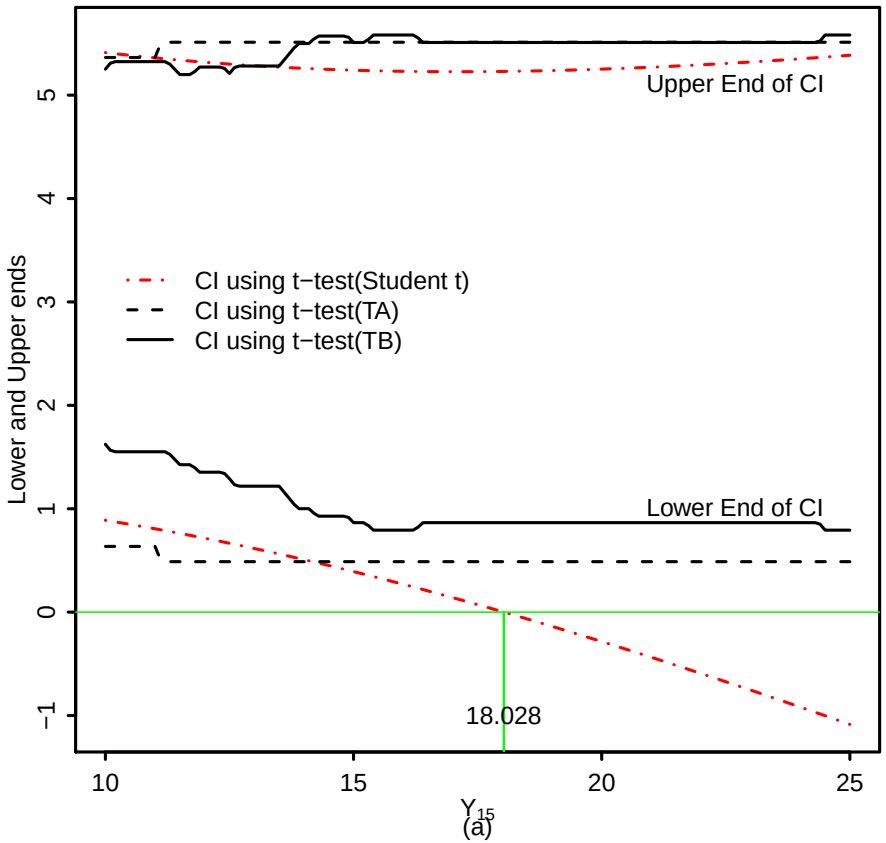
$$T_B = \sqrt{\frac{3n}{2\pi}}\Phi^{-1}\left(\frac{3}{4}\right)\frac{\underset{i\le j}{\text{median}}\left(X_i + X_j\right) - 2\mu}{\underset{i\le j}{\text{median}}\left(|X_i - X_j|\right)} \xrightarrow{d} N(0,1). \tag{2}$$

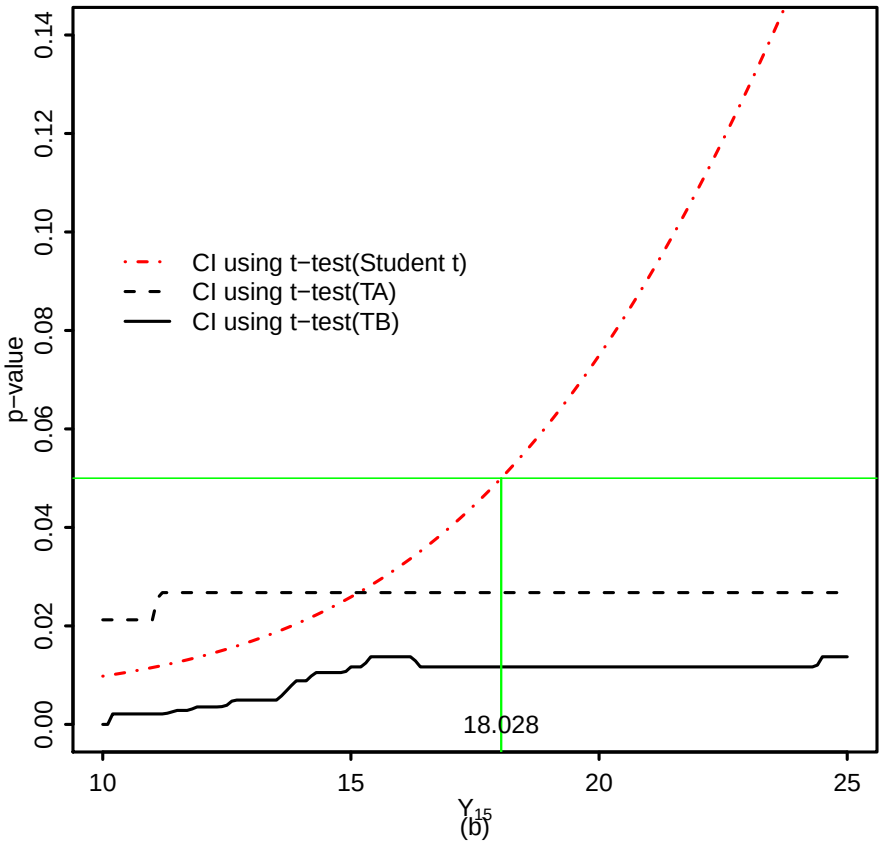
1. Robust statistics와 응용 사례: Robust t-test

- The above Theorems work well with a large sample size because these are based on asymptotic standard normal distribution.
- Recently, Park and Wang (2018) developed the `rt.test` R package used the empirical distributions instead of the asymptotic standard normal distribution. Using the `rt.test`, we can carry out robustified t-test easily.
- Also, we can obtain the confidence intervals using the above robustified test statistics. By checking if zero is included inside each interval, we can test the hypothesis

$$H_0 : \mu_d = 0 \quad \text{and} \quad H_1 : \mu_d \neq 0.$$

1. Robust statistics와 응용 사례: Robust t-test





1. Robust statistics와 응용 사례: Robust t-test

This idea can be easily applied to control charting which is similar to a confidence interval.

- Phase I: use robustified control chart.
- Phase II: use conventional control chart.

This work is partially done (balanced case). See `rcc` function in `rQCC`.

```
> library("rQCC")
> help(rcc)
> tmp = c(
72, 84, 79, 49, 56, 87, 33, 42, 55, 73, 22, 60, 44, 80, 54, 74,
97, 26, 48, 58, 83, 89, 91, 62, 47, 66, 53, 58, 88, 50, 84, 69,
57, 47, 41, 46, 13, 10, 30, 32, 26, 39, 52, 48, 46, 27, 63, 34,
49, 62, 78, 87, 71, 63, 82, 55, 71, 58, 69, 70, 67, 69, 70, 94,
55, 63, 72, 49, 49, 51, 55, 76, 72, 80, 61, 59, 61, 74, 62, 57 )
> data2 = matrix(tmp, ncol=4, byrow=TRUE)
> rcc(data2, loc="HL2", scale="shamos")
     LCL       CL      UCL
36.99703 59.26250 81.52797
```

Robust Design (Dual Response)

- The process mean response function.

$$\hat{M}(\mathbf{x}) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i + \sum_{i=1}^{k} \hat{\beta}_{ii} x_i^2 + \sum_{i<j}^{k} \hat{\beta}_{ij} x_i x_j.$$

- The process variance response function.

$$\hat{V}(\mathbf{x}) = \hat{\eta}_0 + \sum_{i=1}^{k} \hat{\eta}_i x_i + \sum_{i=1}^{k} \hat{\eta}_{ii} x_i^2 + \sum_{i<j}^{k} \hat{\eta}_{ij} x_i x_j.$$

In the above, we need to estimate $M(\mathbf{x})$, $V(\mathbf{x})$, β and η

- The $M(\mathbf{x})$ and $V(\mathbf{x})$ can be estimated using robust estimators such as HL and Shamos estimators. For more details (HL and other estimators), see Talk-2 at ▸ Seminar/2018
- The β and η can be estimated using the regression method.

1. Robust statistics와 응용 사례: Robust Design/강건설계

Refer to Park and Leeds (2016) and Talk-2 at ▸ Seminar/2018

Method A: $\hat{M}(\mathbf{x})$ using the sample **mean** and
$\hat{V}(\mathbf{x})$ using the sample **variance**.
(BASELINE – without contamination!)

Method B: $\hat{M}(\mathbf{x})$: **median** and $\hat{V}(\mathbf{x})$: the squared **MAD**

Method C: $\hat{M}(\mathbf{x})$: **median** and $\hat{V}(\mathbf{x})$: the squared **IQR**

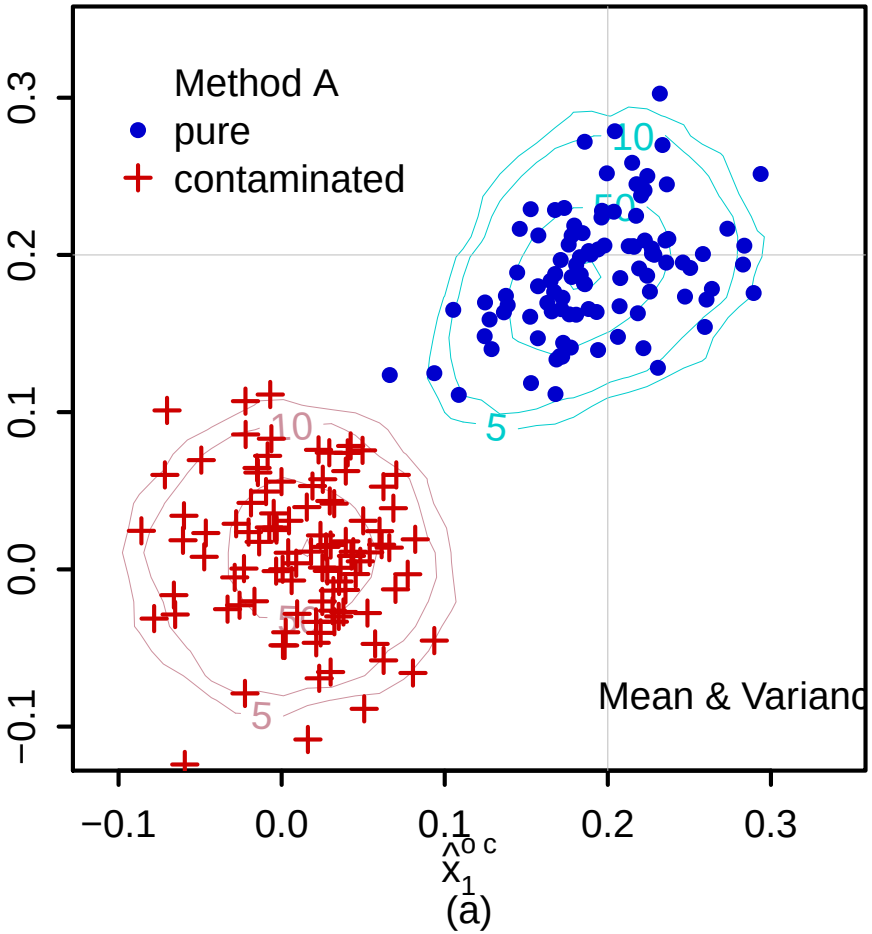
Method D: $\hat{M}(\mathbf{x})$: **HL (Hodges-Lehmann)** and
$\hat{V}(\mathbf{x})$: the squared **Shamos**

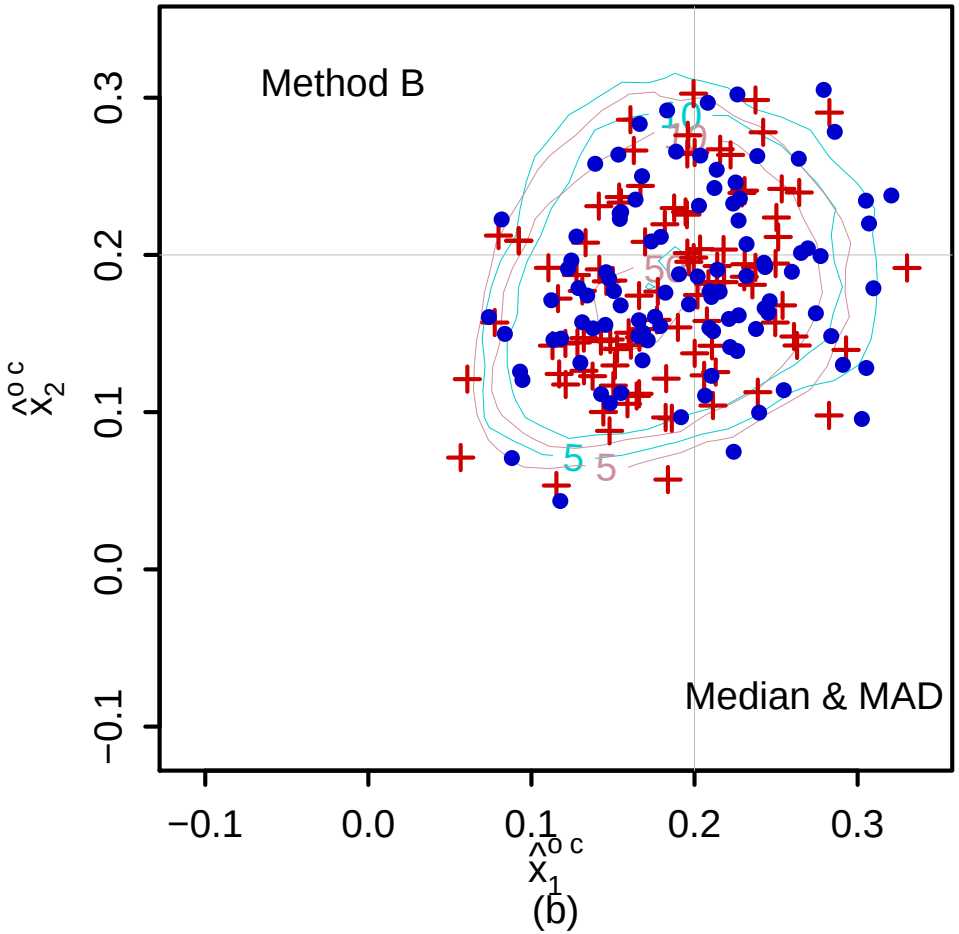
Method E: $\hat{M}(\mathbf{x})$: **median** and $\hat{V}(\mathbf{x})$: the squared **Shamos**

Method F: $\hat{M}(\mathbf{x})$: **HL** and $\hat{V}(\mathbf{x})$: the squared **MAD**.

Method G: $\hat{M}(\mathbf{x})$: **HL** and $\hat{V}(\mathbf{x})$: the squared **IQR**.

1. Robust statistics와 응용 사례: Robust Design/강건설계





Recall: Robust Design (Dual Response)

$$\text{Mean response}: \hat{M}(\mathbf{x}) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i + \sum_{i=1}^{k} \hat{\beta}_{ii} x_i^2 + \sum_{i<j} \hat{\beta}_{ij} x_i x_j$$

$$\text{Variance response}: \hat{V}(\mathbf{x}) = \hat{\eta}_0 + \sum_{i=1}^{k} \hat{\eta}_i x_i + \sum_{i=1}^{k} \hat{\eta}_{ii} x_i^2 + \sum_{i<j} \hat{\eta}_{ij} x_i x_j$$

Unbalanced Data Set

i	x_{i1}	x_{i2}	Y_{ir_i}							$\overline{Y}_i$	S_i^2
1	-1	-1	84.3	57.0	56.5					65.93	253.06
2	0	-1	75.7	87.1	71.8	43.8	51.6			66.00	318.28
3	1	-1	65.9	47.9	63.3					59.03	94.65
4	-1	0	51.0	60.1	69.7	84.8	74.7			68.06	170.35
5	0	0	53.1	36.2	61.8	68.6	63.4	48.6	42.5	53.46	139.89
6	1	0	46.5	65.9	51.8	48.4	64.4			55.40	83.11
7	-1	1	65.7	79.8	79.1					74.87	63.14
8	0	1	54.4	63.8	56.2	48.0	64.5			57.38	47.54
9	1	1	50.7	68.3	62.9					60.63	81.29

Theorem 3

Let $Y_1, \ldots, Y_r$ be a random sample of size r from the probability density function $f(y)$ with a finite fourth moment and let $\mu = E(Y)$ and $\theta_k = E(Y-\mu)^k$, $k = 2, 3, 4$. Then we have

$$\mathrm{Var}(\overline{Y}) = \frac{1}{r}\mu \quad \text{and} \quad \mathrm{Var}(S^2) = \frac{1}{r}(\theta_4 - \frac{r-3}{r-1}\theta_2^2).$$

Especially, if Y_i's have independent and identical normal distribution, then $\mathrm{Var}(S^2) = 2\sigma^4/(r-1)$.

Proof.

See Casella and Berger (2002). □

- $\mathrm{Var}(\overline{Y}) \propto 1/r$.
- $\mathrm{Var}(\overline{S^2}) \propto$?. Under the normality, $\mathrm{Var}(\overline{S^2}) \propto 1/(r-1)$.

2. Missing · Incomplete: RD with unbalanced samples

Under the normality assumption, we can solve this problem by using the weighted least squares (WLS) regression instead of the ordinary least squares (OLS) regression (Cho and Park, 2005).

OLS versus WLS

- OLS: $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ and $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.
- WLS: $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{W}^{-1})$ and $\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$.

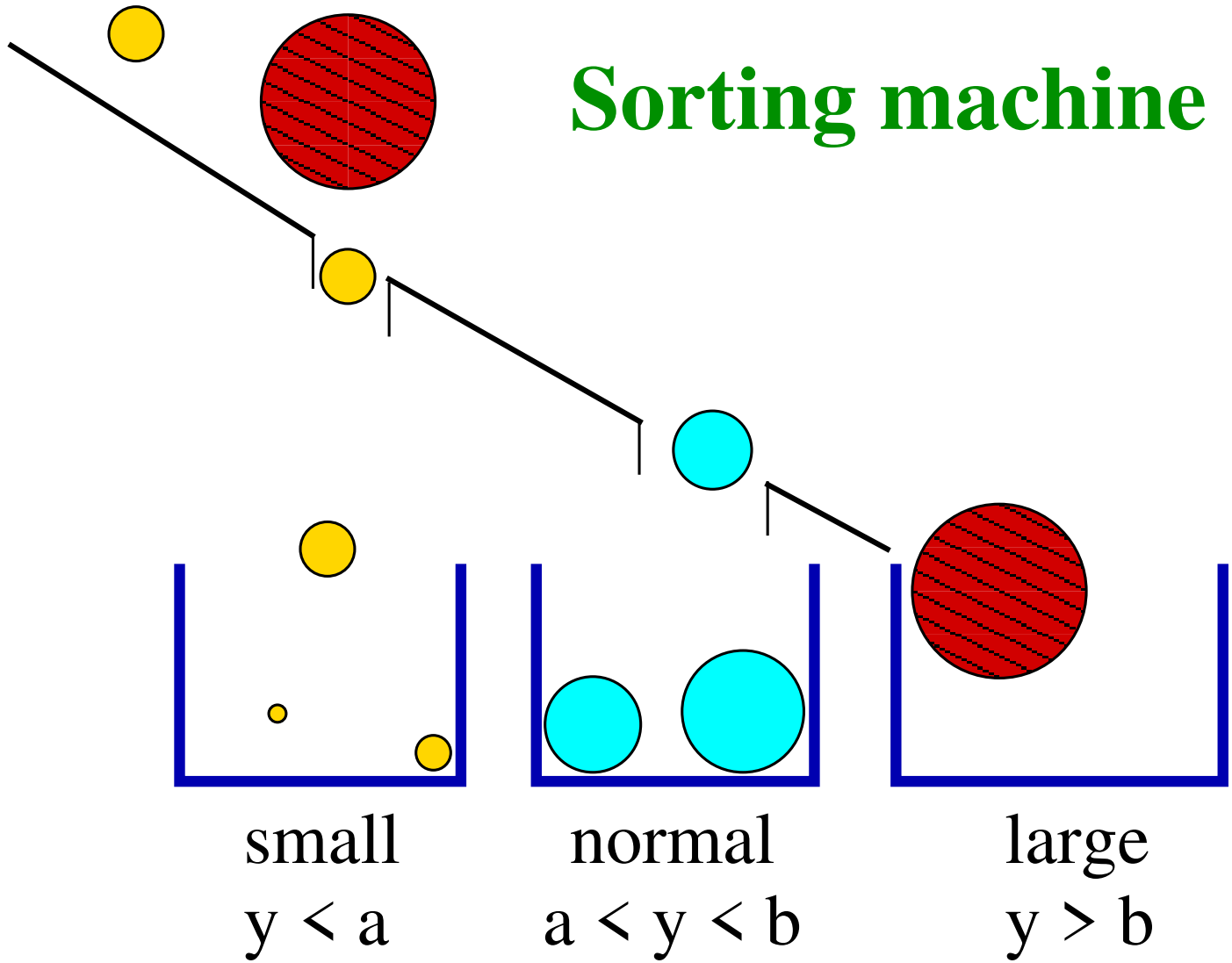
Mean and Variance responses

- Mean response: $\mathbf{W} = \mathrm{diag}(r_1, r_2, \ldots, r_n)$
- Variance response: $\mathbf{W} = \mathrm{diag}(r_1 - 1, r_2 - 1, \ldots, r_n - 1)$

where n is the number of the design points.

What if normality is **not** satisfied

$\mathrm{Var}(\overline{S^2})$ is **not** proportional to $1/(r-1) \Rightarrow$ **multiple imputation**.


Sorting machine
small
y < a
normal
a < y < b
large
y > b

Incomplete Data with grouping

			Full observations					Interval observations		
i	x_{i1}	x_{i2}	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}	$(-\infty, 45)$	$[45, 55]$	$(55, \infty)$
1	-1	-1	55.1	61.4	53.5	72.4	62.6	2	17	81
2	0	-1	65.5	59.2	60.4	57.3	65.0	2	58	40
3	1	-1	58.7	63.3	56.9	49.4	67.7	4	18	78
4	-1	0	61.3	52.1	54.3	47.3	57.9	6	46	48
5	0	0	54.5	47.8	49.8	44.4	51.8	8	84	8
6	1	0	45.0	54.1	62.0	59.0	55.8	6	47	47
7	-1	1	50.5	56.0	54.3	60.2	47.8	5	33	62
8	0	1	52.3	60.5	53.8	62.1	57.9	7	43	50
9	1	1	75.5	44.0	83.7	58.0	56.5	4	23	73

- All observations: Clearly the best
- Full observations only: (measurement cost is expensive).
- Interval observations only: (measurement cost is cheap or free).

Which of full or interval is better? (It depends on the sample size).

Recall: Robust Design (Dual Response)

$$\text{Mean response}: \hat{M}(\mathbf{x}) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i + \sum_{i=1}^{k} \hat{\beta}_{ii} x_i^2 + \sum_{i<j}^{k} \hat{\beta}_{ij} x_i x_j$$

$$\text{Variance response}: \hat{V}(\mathbf{x}) = \hat{\eta}_0 + \sum_{i=1}^{k} \hat{\eta}_i x_i + \sum_{i=1}^{k} \hat{\eta}_{ii} x_i^2 + \sum_{i<j}^{k} \hat{\eta}_{ij} x_i x_j$$

For the application of incomplete/grouped data to the robust design, see Lee and Park (2006). (Below, $n = 5$ full obs. and 100 interval obs.)

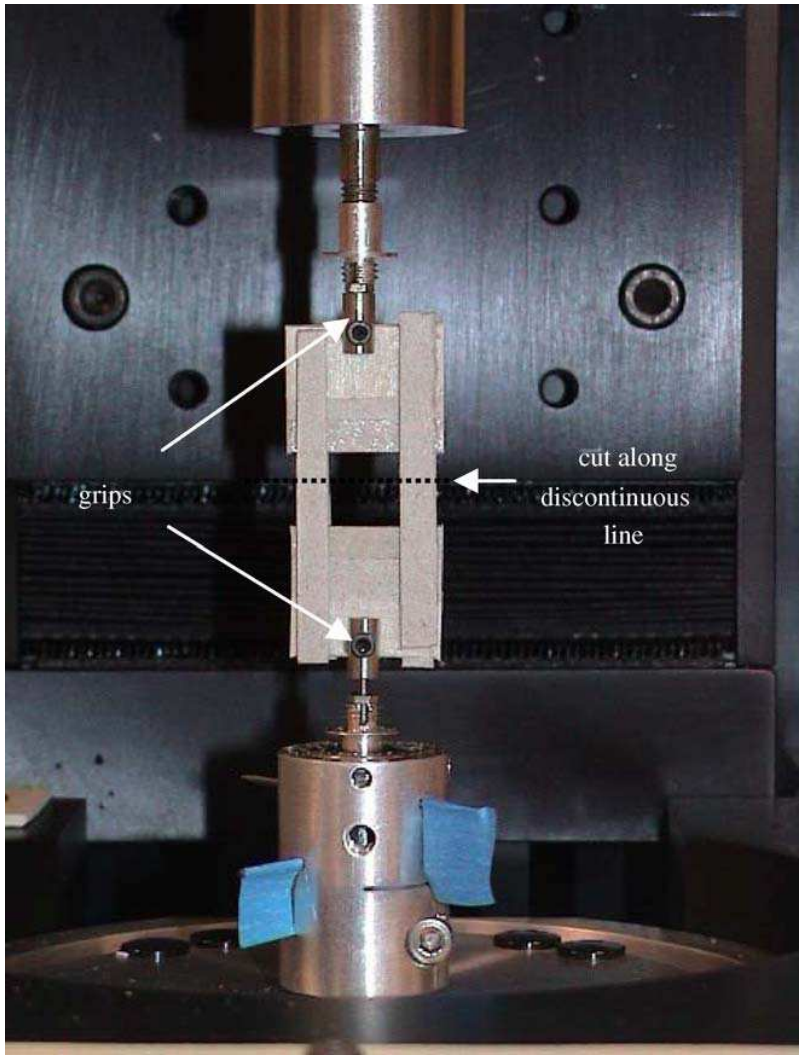
Empirical bias and MSE of the mean and variance under considered method

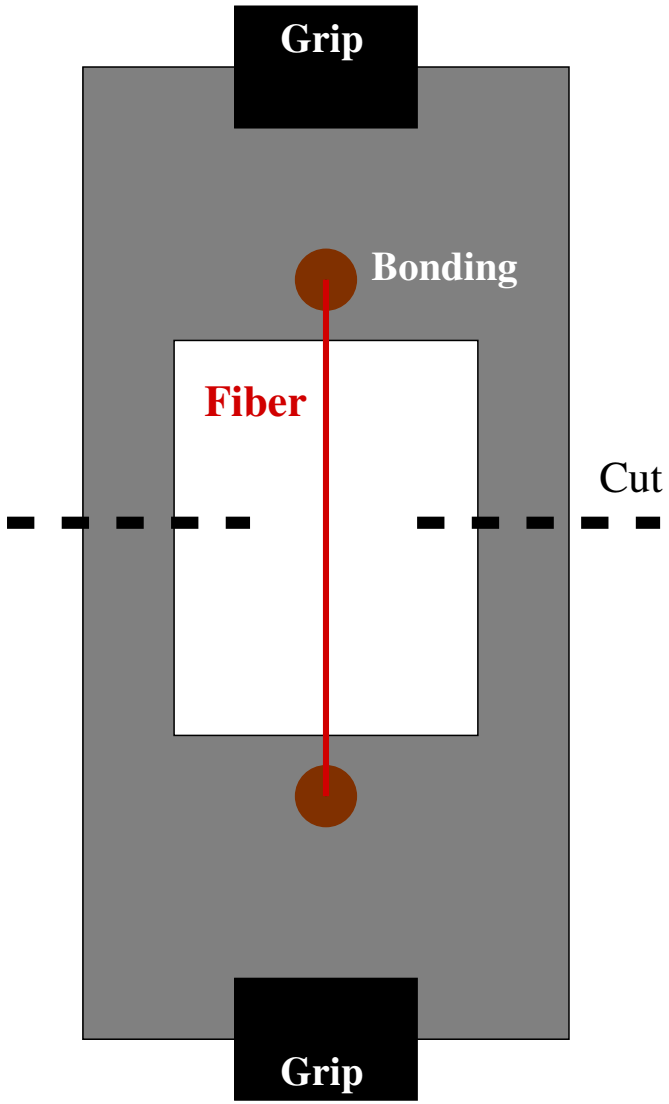
	Full obs. only		All observations		Interval obs. only	
Estimate	$\hat{\mu}_1$	$\hat{\sigma}_1^2$	$\hat{\mu}_2$	$\hat{\sigma}_2^2$	$\hat{\mu}_3$	$\hat{\sigma}_3^2$
Bias	0.0956	−0.9653	0.0059	2.8018	0.0009	4.5426
Variance	20.2007	4777.45	1.2125	762.0400	1.3103	989.9277
MSE	20.2098	4778.38	1.2125	769.8901	1.3103	1010.5630

2. Missing · Incomplete: RD with incomplete/grouped data

- Full observations are costly.
  Interval observatios are cheap or free.
- Curvature of profile likelihood can be used for precision of estimators, which can be used for equivalent sample sizes with the same precision (Lee and Park, 2006).
- The parameter estimation with incomplete/grouped data is tricky.
  The EM, MCEM, QEM can be used (Park, 2018b).
  The MI can also be used but this is also tricky.
  In general, the model is known, the MLE (EM) is better than the MI.
- This idea can be applied to various applications with parameter estimates.

Illustrative Example: Tensile Testing Equipment





2. Missing · Incomplete: Competing risks

Most multi-modal strength analyses of materials have been studied based on the so-called weakest link theory which requires two assumptions (Beetz, 1982; Goda and Fukunaga, 1986):

Assumptions

A1 The material contains inherently many strength-limiting defects, and its strength depends on the weakest defect of all of them.

A2 There are no interactions among the defects.

What if the above assumptions are **not** satisfied

multiple imputation can be considered.

Strength data with three fracture causes (modes)

Strength	Mode	Strength	Mode	Str	Mode	Str	Mode
54	{3}	7	{1, 2, 3}	86	{2}	104	{1}
143	{2}	81	{3}	141	{1}	89	{3}
97	{3}	52	{3}	79	{3}	9	{3}
104	{3}	40	{3}	23	{3}	111	{1, 2, 3}
71	{1, 2}	82	{2}	8	{3}	150	0
98	{1}	3	{3}	17	{3}	79	{2}
24	{2}	130	{2}	41	{2}	94	{2}
138	{3}	5	{3}	43	{2, 3}	150	0
38	{3}	32	{2}	9	{3}	77	{2}
78	{3}	16	{3}	92	{2}	76	{3}
150	0	33	{3}	80	{2}	100	{2}
46	{3}	137	{1, 2}	92	{3}	108	{2}
109	{1}	71	{1}	60	{2}	88	{1}
7	{3}	11	{3}	150	0	150	0
42	{2}	6	{3}	43	{3}	124	{1, 2}

Specimens in tensile strength experiments are broken down due to

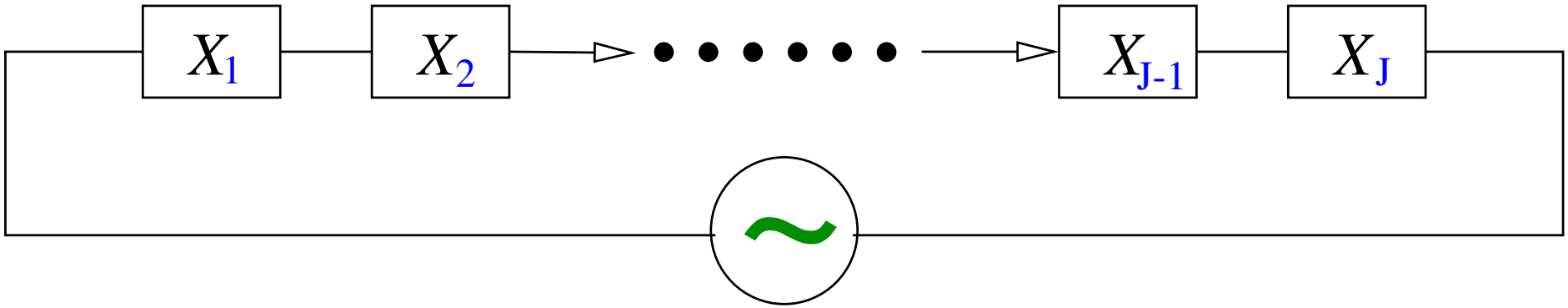
- several causes (**competing risks**)
- with the cause of fracture not properly identified (**missing**)
- along with censoring due to time and cost considerations on experiments.

For example, the fracture causes are due to:

- a surface defect (**mode 1**),
- an inner defect (**mode 2**), and
- an end effect at the clamp to hold the specimen (**mode 3**).
- The censored observations are denoted by **0**.
  (The observations were censored at **150** in the previous data set).

NOTE: In competing risks literature, **missing cause** is called **masking**.

The compering risks can be modeled as a system in series.



Let $X = \min(X_1, X_2, \ldots, X_J)$. Then the cdf of X is easily obtained as

$$\begin{aligned} F(x|\boldsymbol{\Theta}) &= 1 - P[X > x] = 1 - P[\min(X_1, X_2, \ldots, X_J) > x] \\ &= 1 - P[X_1 > x, X_2 > x, \ldots, X_J > x] \\ &= 1 - P[X_1 > x] \cdot P[X_2 > x] \cdot \cdots \cdot P[X_J > x] \\ &= 1 - \prod_{j=1}^{J} \left\{ 1 - F_j(x|\boldsymbol{\theta}_j) \right\}, \end{aligned}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_J)$.

2. Missing · Incomplete: Load-Sharing

.....

Will be added

....

.....

Will be added

....

3. Future work: Missing with contamination

XXXX

- XXX
- XXX

References

Basu, A., Shioya, H., and Park, C. (2011). Statistical Inference: The Minimum Distance Approach. Monographs on Statistics and Applied Probability. Chapman & Hall.

Beetz, C. P. (1982). The analysis of carbon fibre strength distributions exhibiting multiple modes of failure. Fibre Science Technology, 16:45–59.

Casella, G. and Berger, R. L. (2002). Statistical Inference. Duxbury, Pacific Grove, CA, second edition.

Cho, B.-R. and Park, C. (2005). Robust design modeling and optimization with unbalanced data. Computers & Industrial Engineering, 48:173–180.

Darwin, C. (1876). The Effect of Cross- and Self-fertilization in the Vegetable Kingdom. John Murry, London, 2nd edition.

References

Friendly, M., Dray, S., Wickham, H., Hanley, J., Murphy, D., and Li, P. (2018). `HistData`: Data sets from the history of statistics and data visualization. `https://CRAN.R-project.org/package=HistData`. R package version 0.8-4.

Goda, K. and Fukunaga, H. (1986). The evaluation of the strength distribution of silicon carbide and alumina fibres by a multi-modal Weibull distribution. Journal of Materials Science, 21:4475–4480.

Jeong, R., Son, S. B., Lee, H. J., and Kim, H. (2018). On the robustification of the z-test statistic. Presented at KIIE Conference, Gyeongju, Korea. April 6, 2018.

Lee, S. B. and Park, C. (2006). Development of robust design optimization using incomplete data. Computers & Industrial Engineering, 50:345–356.

Park, C. (2018a). Note on the robustification of the Student t-test statistic using the median and the median absolute deviation. `https://arxiv.org/abs/1805.12256`. ArXiv e-prints.

References

Park, C. (2018b). A quantile variant of the Expectation-Maximization algorithm and its application to parameter estimation with interval data. Journal of Algorithms & Computational Technology, 12:253–272.

Park, C. and Leeds, M. (2016). A highly efficient robust design under data contamination. Computers & Industrial Engineering, 93:131–142.

Park, C. and Wang, M. (2018). Empirical distributions of the robustified t-test statistics. `https://arxiv.org/abs/1807.02215`. ArXiv e-prints.