Efficient Robust Methods and Their Applications

Chanseok Park

Applied Statistics Laboratory
Department of Industrial Engineering
Pusan National University

July 13, 2018

Hosted by SEC

Overview

Overview

1. Robust Parameter Estimation
2. Application I (Robust Design with Contaminated Data)
3. Application II (Robust Design with Model Departure)
4. Application IV (Robustifed t-test under Contamination)
5. R Data Example with `rt.test` package

Overview

1. Robust Parameter Estimation
2. Application I (Robust Design with Contaminated Data)
3. Application II (Robust Design with Model Departure)
4. Application IV (Robustifed t-test under Contamination)
5. R Data Example with `rt.test` package

Overview

Overview

1. Robust Parameter Estimation
2. Application I (Robust Design with Contaminated Data)
3. Application II (Robust Design with Model Departure)
4. Application IV (Robustifed t-test under Contamination)
5. R Data Example with `rt.test` package

Robust Parameter Estimation

Sample mean and variance

$$\overline{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j$$

$$S^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2.$$

- The sample mean and the sample variance are the most widely used estimators for location (mean) and squared scale (variance).
- The problem is that they are very sensitive to contamination.

Example

- Data: $Y = (-2, -1, 0, 1, 2)$
  Mean response: mean $=$ 0 and median $=$ 0
  Var. response: var $=$ 2.5 and IQR $=$ 2

- Data: $(-2, -1, 0, 1, 102)$
  Mean response: mean $=$ 20 and **median = 0**
  Var. response: var $=$ 2102.5 and **IQR = 2**

- In the above, **median** (alternative to mean) and **IQR** (alternative to standard dev.) are illustrated.
- But, there are several other alternatives to them.

Alternative to mean and variance (or, standard deviation)

- Mean: median, Hodge-Lehmann

$$\mathrm{HL} = \underset{i \le j}{\mathrm{median}} \left(\frac{Y_i + Y_j}{2} \right).$$

- Std Deviation: MAD, IQR, Shamos (needs normal-consistency correction).

$$\mathrm{MAD} = \underset{1 \le i \le m}{\mathrm{median}} \left\{ |Y_i - \mathrm{median}(Y)| \right\}$$

$$\mathrm{IQR} = Y_{[3m/4]} - Y_{[m/4]}$$

$$\mathrm{Shamos} = \underset{i \le j}{\mathrm{median}} \left(|Y_i - Y_j| \right)$$

- Then which one should be selected?
- We need to consider (i) **breakdown point** and (ii) **efficiency**.

Robust Parameter Estimation

- **Breakdown point**: the proportion of incorrect observations (e.g. arbitrarily large observations) an estimator can handle.
- **ARE** (asymptotic relative efficiency) is the ratio of variance of MLE to variance of the corresponding estimator.

Properties of Location and Scale Estimators

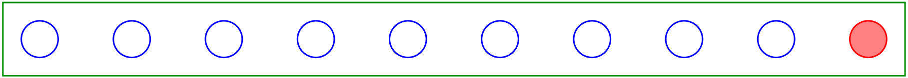
Location	Mean	Median	**Hodges-Lehmann**
Breakdown	0%	**50%**	29%
ARE	**100%**	64%	**96%**

Scale	SD	IQR	MAD	**Shamos**
Breakdown	0%	25%	**50%**	29%
ARE	**100%**	38%	37%	**86%**

- Mean and SD are the MLEs for the location and scale under the normal assumption.

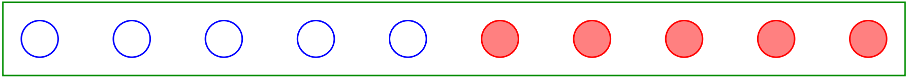
- Mean starts to break down even with a single extreme value (say, ∞).

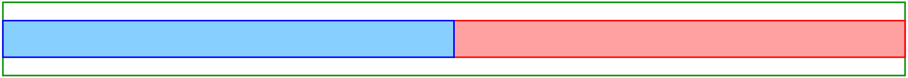
$$\text{Mean} = \frac{1}{10}Y_1 + \frac{1}{10}Y_2 + \cdots + \frac{1}{10}Y_{10} \qquad \text{(0\% breakdown)}$$



- Median starts to break down with **five** or more extremes out of **ten**.

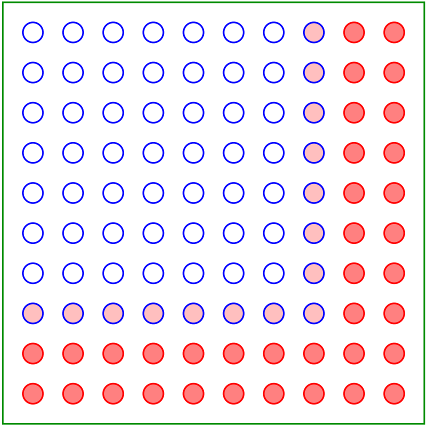
$$\text{Median} = (Y_{(5)} + Y_{(6)})/2 \qquad \text{(50\% breakdown)}$$

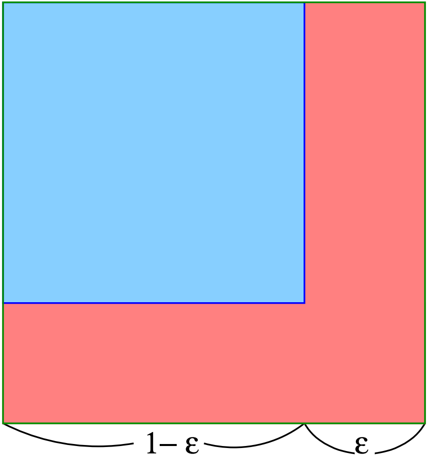




What is the breakdown point of the HL, $\mathrm{median}_{i\le j}(Y_i + Y_j)/2$?

- When 2 out of 10 are contaminated, HL is OK.
  But, when 3 are contaminated, it breaks down.
- Thus, the breakdown point is between 2/10 and 3/10.
- Solving $(1-\epsilon)^2 = 1/2$ for ϵ, we have $\epsilon = 1-(1/2)^{1/2}$.





Robust Parameter Estimation (Breakdown Point)

What is the breakdown point of $\mathrm{median}_{i \le j \le k}(Y_i + Y_j + Y_k)/3$**?**
Solving the following for ϵ ($d = 3$)

$$(1-\epsilon)^d = \frac{1}{2}$$

we have

$$\epsilon = 1 - \left(\frac{1}{2}\right)^{1/d}.$$

$d = 1$ Median	$d = 2$ Hodges-Lehmann	$d = 3$?	d (vary large) close to Mean
0.5 (50%)	0.293 (29.3%)	0.206 (20.6%)	close to 0

It looks like that the median is the best?
But, we also consider the ARE.

Robust Parameter Estimation (ARE)

The RE (relative efficiency) and ARE (asymptotic relative efficiency) are defined as

$$\mathrm{RE}(\hat{\theta}_2, \hat{\theta}_1) = \frac{\mathrm{Var}(\hat{\theta}_1)}{\mathrm{Var}(\hat{\theta}_2)} \times 100\%$$

$$\mathrm{ARE}(\hat{\theta}_2, \hat{\theta}_1) = \frac{\mathrm{AVar}(\hat{\theta}_1)}{\mathrm{AVar}(\hat{\theta}_2)} \times 100\%,$$

where $\hat{\theta}_1$ is a reference or baseline estimator (say, MLE without contamination).

- It is quite difficult to obtain the RE and ARE theoretically (Serfling, 2011).
- One can use Monte Carlo simulation.

Robust Parameter Estimation

Recall: Properties of Location and Scale Estimators

Location	Mean	Median	**Hodges-Lehmann**
Breakdown	0%	**50%**	29%
ARE	**100%**	64%	**96%**

Scale	SD	IQR	MAD	**Shamos**
Breakdown	0%	25%	**50%**	29%
ARE	**100%**	38%	37%	**86%**

Application I (Robust Design with Contaminated Data)

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.

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

Note: original x are centered and re-scaled to $x \in [-1, 1]$.
We need to estimate, $M(\mathbf{x})$, $V(\mathbf{x})$, β and η.
The β and η can be estimated using the least squares method, etc.

Application I (Robust Design with Contaminated Data)

Refer to Park and Leeds (2016).

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**.

Application I (Robust Design with Contaminated Data)

Data Description

We will use a case study from Park (2013) in order to evaluate the outlier-resistance properties of Methods A–G.

- A company produces multi-filament microfiber tows.
- Control factors: polymer **temperature ($x_{1,i}$)** and polymer feeding **speed ($x_{2,i}$)**.
- The diameter (Y): the main quality issue.
  Its nominal **target value**: $T_0 = 50$ microns.
- The 3×3 factorial design ($i = 1, 2, \ldots, 9$). We observe the diameters of 10 fibers ($j = 1, 2, \ldots, 10$).
- The original covariates have been centered and re-scaled so that $x_{1,i}$ and $x_{2,i}$ are in $[-1, 1]$.

Application I (Robust Design with Contaminated Data)

The original observation ($Y_{11} = 73.94$) will be modified later.

i	$x_{1,i}$	$x_{2,i}$	Y_{ij}				
1	-1	-1	$\boxed{Y_{11}}$	76.09	73.39	79.82	76.47
			73.43	76.89	77.55	77.12	74.79
2	0	-1	67.30	64.55	62.08	58.18	66.36
			63.49	63.56	65.91	65.61	65.05
3	1	-1	94.03	93.67	91.80	86.34	93.24
			91.45	91.19	87.71	90.33	92.71
4	-1	0	66.93	63.35	64.55	63.47	60.23
			62.58	62.63	63.45	66.29	65.47
5	0	0	51.23	51.03	53.16	52.84	50.06
			50.02	52.42	53.32	51.35	53.57
6	1	0	80.58	78.10	80.44	76.83	83.11
			84.45	78.70	77.04	81.00	79.27
			......	......	......	......	......

Application I (Robust Design with Contaminated Data)

The results with the original observation ($Y_{11} = 73.94$)

$$\hat{M}(\mathbf{x}) = 51.741 + 7.750x_1 + 8.053x_2 + 20.262x_1^2 + 19.939x_2^2 - 0.038x_1x_2.$$
$$\hat{V}(\mathbf{x}) = 0.841 - 0.015x_1 - 0.068x_2 + 0.620x_1^2 + 0.421x_2^2 - 0.339x_1x_2.$$

Then, by minimizing

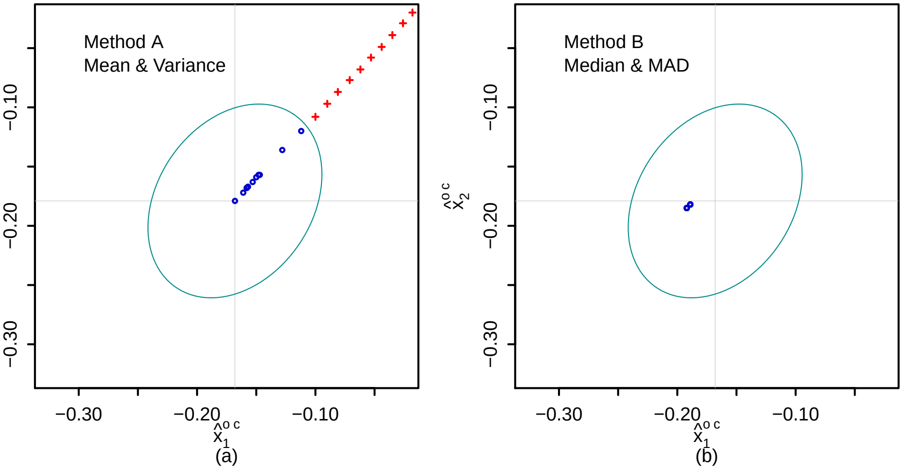
$$\{\hat{M}(\mathbf{x}) - 50\}^2 + \exp\left(\hat{V}(\mathbf{x})\right)$$

subject to $|x_1| \le 1$ and $|x_2| \le 1$, the optimal operating conditions were given by

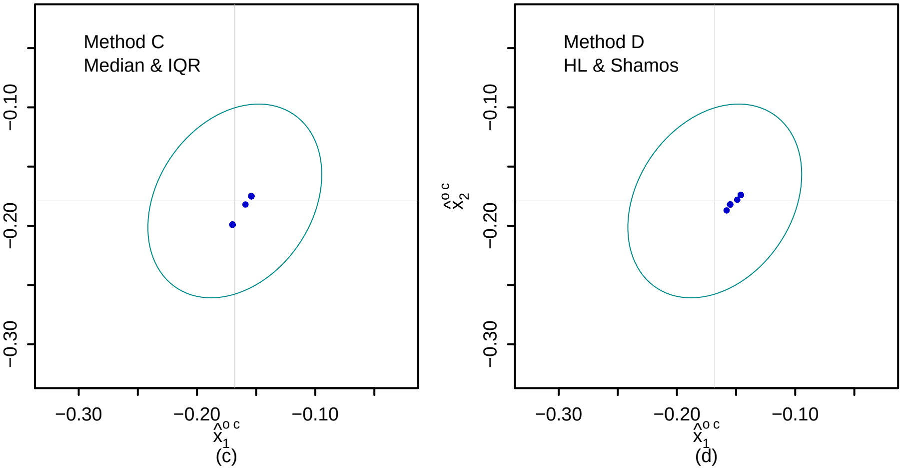
$$\hat{\mathbf{x}}^{\mathrm{oc}} = (\hat{x}_1^{\mathrm{oc}}, \hat{x}_2^{\mathrm{oc}}) = (-0.168, -0.179).$$

Next, we will do the above analysis again with contaminated data sets, where $\boxed{Y_{11}}$ is changed to $10, 20, 30, \ldots, 200$.

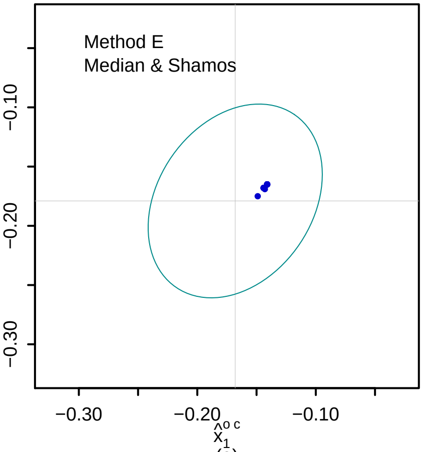
Application I (Robust Design with Contaminated Data)



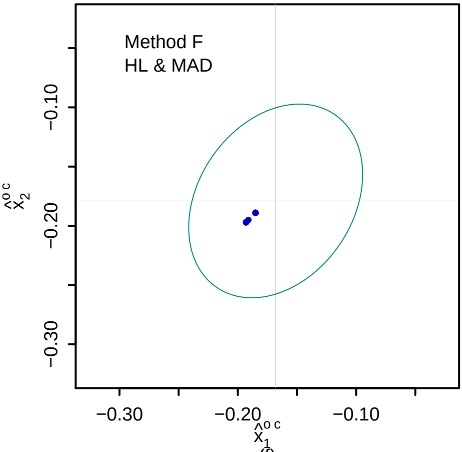
Application I (Robust Design with Contaminated Data)



Application I (Robust Design with Contaminated Data)



(e)



(f)

Application I (Robust Design with Contaminated Data)

- **Euclidean distance**:

$$d = \sqrt{\left\{\hat{x}^{\mathrm{oc}}_{1,\mathrm{noise}} - (-0.168)\right\}^2 + \left\{\hat{x}^{\mathrm{oc}}_{2,\mathrm{noise}} - (-0.179)\right\}^2}.$$

  - It is *not* invariant w.r.t. scale.
  - The statistical distribution is unknown and this makes it difficult to measure the discrepancy statistically.

- **Mahalanobis distance**:

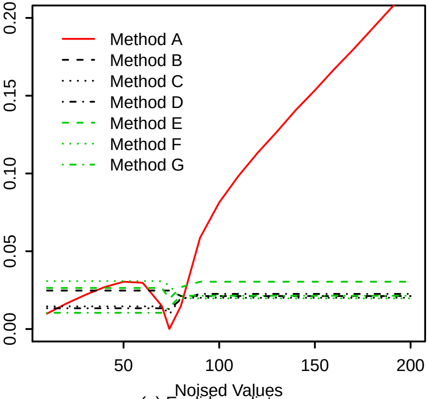
$$D_M = (\hat{\mathbf{x}}^{\mathrm{oc}}_{\mathrm{noise}} - \hat{\mathbf{x}}^{\mathrm{oc}})'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\mathbf{x}}^{\mathrm{oc}}_{\mathrm{noise}} - \hat{\mathbf{x}}^{\mathrm{oc}}), \tag{1}$$

where $\hat{\mathbf{x}}^{\mathrm{oc}} = (-0.168, -0.179)$. Note the covariance matrix $\hat{\boldsymbol{\Sigma}}$ was estimated using the bootstrapping method proposed by Park (2013). Taking the inverse of $\hat{\boldsymbol{\Sigma}}$ resulted in

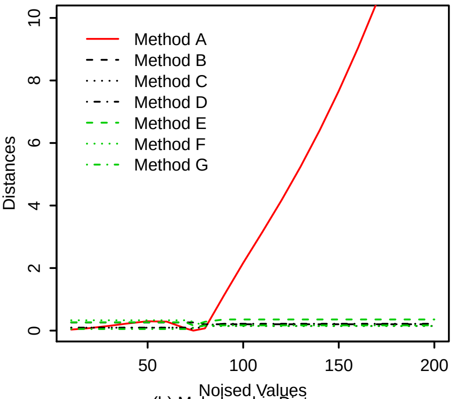
$$\hat{\boldsymbol{\Sigma}}^{-1} = \begin{bmatrix} 498.0663 & -122.6756 \\ -122.6756 & 402.3380 \end{bmatrix}.$$

  - It is computationally costly to calculate $\hat{\boldsymbol{\Sigma}}^{-1}$.

Application I (Robust Design with Contaminated Data)



(a) Euclidean distances



(b) Mahalanobis Distances

Application I (Robust Design with Contaminated Data)

Simulation Studies

- We make the assumption that the true mean process $M(\mathbf{x})$ and variance process $V(\mathbf{x})$ are known to be the following:

  $$M(\mathbf{x}) = T_0 + 5(x_1^2 + x_2^2) \text{ and } V(\mathbf{x}) = 1 + (x_1 - 1)^2 + (x_2 - 1)^2,$$

  where the target $T_0 = 50$. At each design point i,

  $$Y_{ij} \sim N\Big(M(\mathbf{x}_i), V(\mathbf{x}_i)\Big),$$

  where $\mathbf{x}_i = (x_{1i}, x_{2i})$ with $x_{1,i} = -1, 0, 1$ and $x_{2,i} = -1, 0, 1$, and $i = 1, 2, \ldots, 9$ (3×3 design), and $j = 1, 2, \ldots, 50$.
- Then we contaminate 5 of each of the original $m = 50$ uncontaminated responses (10% contamination) from the first simulation. This was done by randomly adding 100 to the originally uncontaminated value.

Application I (Robust Design with Contaminated Data)

Theoretical optimal conditions with pure data

Denoting the MSE (squared loss) as $\phi(\mathbf{x})$, we have

$$\begin{aligned}\phi(\mathbf{x}) &= \left\{M(\mathbf{x}) - T_0\right\}^2 + V(\mathbf{x}) \\ &= 25(x_1^2 + x_2^2)^2 + 1 + (x_1 - 1)^2 + (x_2 - 1)^2\end{aligned}$$

By setting $\partial\phi/\partial x_1 = 0$ and $\partial\phi/\partial x_2 = 0$, it is immediate that

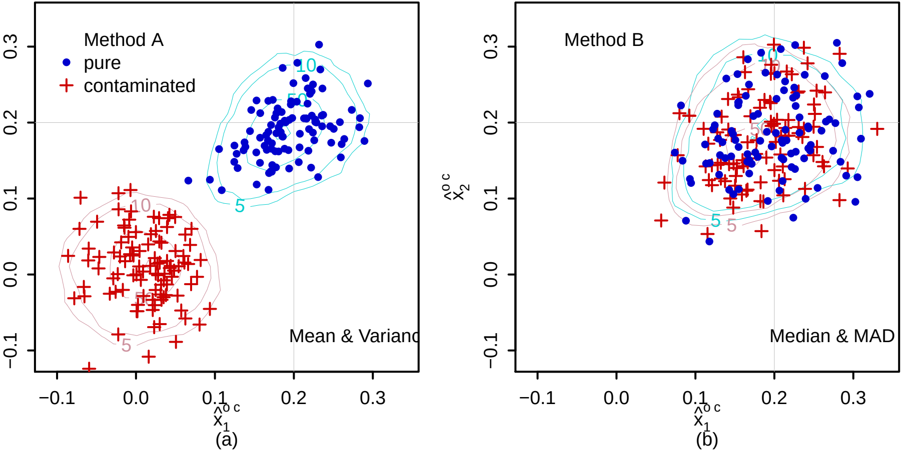
$$\frac{\partial\phi}{\partial x_1} = 100(x_1^2 + x_2^2)x_1 + 2(x_1 - 1) = 0 \tag{2}$$

$$\frac{\partial\phi}{\partial x_2} = 100(x_1^2 + x_2^2)x_2 + 2(x_2 - 1) = 0 \tag{3}$$

Solving the above, we have

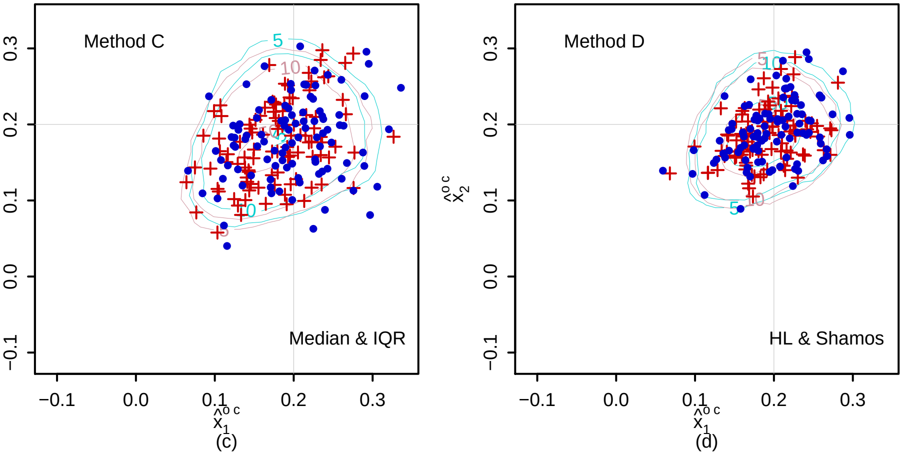
$$(x_1^{\mathrm{oc}}, x_2^{\mathrm{oc}}) = (0.2, 0.2)$$

Application I (Robust Design with Contaminated Data)

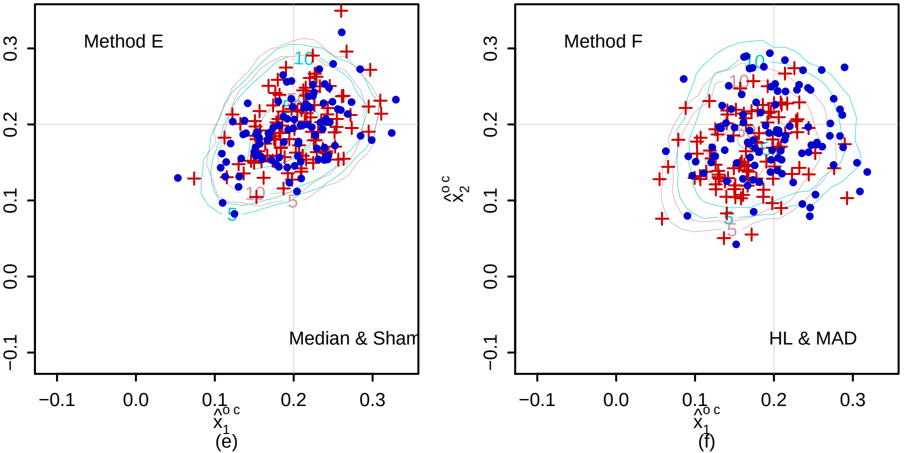


Note: We plot only 100 points from 10,000 simulation results.

Application I (Robust Design with Contaminated Data)



Application I (Robust Design with Contaminated Data)



Application I (Robust Design with Contaminated Data)

A well known and useful comparison of two estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, is obtained by a direct comparison of their variances (see Section 2.2 of Lehmann (1999)) and is termed the relative efficiency of $\hat{\theta}_2$ to $\hat{\theta}_1$:

$$\mathrm{RE}(\hat{\theta}_2, \hat{\theta}_1) = \frac{\mathrm{Var}(\hat{\theta}_1)}{\mathrm{Var}(\hat{\theta}_2)}.$$

However, the relative efficiency compares variances and, in our framework we are considering **two-dimensional** estimator (namely, location and scale at the same time). Therefore, the usual concept of efficiency is not directly applicable. In order to deal with this issue,

We propose the generalized variance using the determinant instead of the conventional variance. For more details, see Lee and Park (2017). Then the relative efficiency of Method 2 to Method 1 in our robust design framework is

$$\mathrm{RE}(\mathrm{Method}\ 2, \mathrm{Method}\ 1) = \frac{\mathrm{gVar}(\mathrm{Method}\ 1)}{\mathrm{gVar}(\mathrm{Method}\ 2)}.$$

Application I (Robust Design with Contaminated Data)

Table 1: Relative efficiencies of each method to Method A without contamination based on the generalized variance.

Method	A	B	C	D	E	F	G
No contamination	**100.0**%	25.1%	25.2%	**82.9**%	59.6%	32.2%	32.2%
Contamination	**1.9**%	25.6%	25.6%	**87.6**%	68.7%	26.5%	26.3%

Application II (Robust Design with Model Departure)

Robust Design with Normal Model Departure

- Similar to Application I.
- We considered **normal model departure** instead of contamination.
- Refer to "**Park**, Ouyang, **Byun** and Leeds (2017)."

We briefly introduce their results.

Application II (Robust Design with Model Departure)

Recall Dual Response Model

- 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.

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

Application II (Robust Design with Model Departure)

Recall Simulation Studies of Application I

Assume $M(\mathbf{x})$ and $V(\mathbf{x})$ are known as

$$M(\mathbf{x}) = T_0 + 5(x_1^2 + x_2^2) \text{ and } V(\mathbf{x}) = 1 + (x_1 - 1)^2 + (x_2 - 1)^2,$$

where the target $T_0 = 50$. At each design point i and $\mathbf{x}_i = (x_{1i}, x_{2i})$ $(j = 1, 2, \ldots, 50)$,

$$Y_{ij} \sim N\big(M(\mathbf{x}_i), V(\mathbf{x}_i)\big),$$

which is equivalent to

$$Y_{ij} = M(\mathbf{x}_i) + U_i, \quad \text{where} \quad U_i \sim N\big(0, V(\mathbf{x}_i)\big).$$

Question: what if U_i is **not** normal?

Application II (Robust Design with Model Departure)

We omit the index i for brevity.

- Normal: $Y = M(\mathbf{x}_i) + U_i$
  where $U \sim N\big(0, V(\mathbf{x})\big)$
- Uniform: $Y = M(\mathbf{x}_i) + U_i$
  where $U \sim \mathrm{Uniform}(-\sqrt{3V(\mathbf{x})}, \sqrt{3V(\mathbf{x})})$.
- Logistic: $Y = M(\mathbf{x}_i) + \sqrt{3}/\pi \cdot U$
  where U is from $\mathrm{Logistic}(0, V(\mathbf{x})^{1/2})$.
- Laplace (or double exponential): $Y = M(\mathbf{x}) + U/\sqrt{2}$
  where U is from $\mathrm{Laplace}(0, V(\mathbf{x})^{1/2})$.
- Student t distribution: $Y = M(\mathbf{x}) + (\nu - 2)/\nu \cdot V(\mathbf{x})^{1/2} \cdot U$,
  where U is from the t-distribution with ν degrees of freedom.

Consequently, in all the model departure scenarios above, we have the same $E(Y)$ and $\mathrm{Var}(Y)$ as

$$E(Y) = M(\mathbf{x}) \quad \text{and} \quad \mathrm{Var}(Y) = V(\mathbf{x}).$$

Application II (Robust Design with Model Departure)

As before, the relative efficiency of Method 2 to Method 1 is used

$$\mathrm{RE}(\mathrm{Method\ 2}, \mathrm{Method\ 1}) = \frac{\mathrm{gVar}(\mathrm{Method\ 1})}{\mathrm{gVar}(\mathrm{Method\ 2})}.$$

It should be noted that the generalized variance (gVar) is investigated further in Lee and Park (2017) where the generalized mean square error is defined. Using this approach, we can also define the **generalized bias**.

Application II (Robust Design with Model Departure)

We used the same methods as we did in Application I. We recall

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**.

Relative efficiencies (percent) of each method under consideration

Table 2: Relative efficiencies (percent) of each method under consideration to Method A based on the generalized variance. The kurtosis of each distribution is shown in the last column.
(The kurtosis vs. model departure can be an interesting future topic).

Underlying distribution	Method A	B	C	D	E	F	G	Kurtosis $(\kappa - 3)$
Normal	100.0	24.1	23.9	82.5	59.1	30.9	30.5	0
Uniform	100.0	7.4	7.6	80.4	26.2	14.5	14.7	-1.2
Logistic	100.0	34.9	34.7	101.4	83.3	40.2	39.9	1.2
Laplace	100.0	37.3	37.4	98.5	103.7	36.0	36.2	3
t (df=5)	100.0	57.8	57.2	151.2	131.7	63.9	63.2	6
t (df=4)	100.0	83.2	83.0	200.9	180.4	90.3	90.1	∞
t (df=3)	100.0	164.5	164.3	333.8	323.3	167.5	168.0	∞

We will shows the above results using the bar-plot.

Application II (Robust Design with Model Departure)

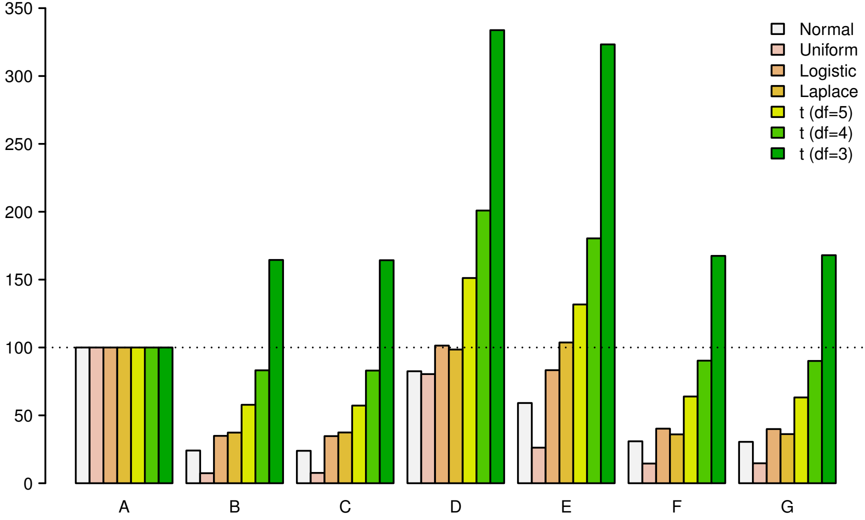


Figure 1: Relative efficiencies based on Table 35.

Application III (Robustified t-test under Contamination)

Basic Idea

Recall

$$T = \frac{\boxed{\bar{X}} - \mu}{\boxed{S}/\sqrt{n}} \tag{4}$$

Thus, one can use the median (or Hodges-Lehmann) instead of $\boxed{\bar{X}}$ and the MAD (or Shamos) instead of $\boxed{S}$ in the above.

- **Is it enough?**
- Then **is it distributed as** $N(0,1)$ **or** t**-distribution?**

Application III (Robustified t-test under Contamination)

Theorem 1 (Pivot with median and MAD)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a location-scale family with location μ and scale σ. Then the statistic below is a **pivotal** *quantity:*

$$\frac{\operatorname*{median}_{1\le i\le n} X_i - \mu}{\operatorname*{MAD}_{1\le i\le n} X_i/\sqrt{n}} = \frac{\operatorname*{median}_{1\le i\le n} X_i - \mu}{\operatorname*{median}_{1\le i\le n} \left| X_i - \operatorname*{median}_{1\le i\le n} X_i \right| / \sqrt{n}} \tag{5}$$

Proof.

See Park (2018) and Jeong et al. (2018). □

Application III (Robustified t-test under Contamination)

Theorem 2 (Asymptotic Normality with median and MAD)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$. Then we have

$$\sqrt{\frac{2n}{\pi}}\Phi^{-1}\left(\frac{3}{4}\right) \cdot \frac{\operatorname*{median}_{1\le i\le n} X_i - \mu}{\operatorname*{median}_{1\le i\le n} \left| X_i - \operatorname*{median}_{1\le i\le n} X_i \right|} \xrightarrow{d} N(0,1).$$

Proof.

See Park (2018) and Jeong et al. (2018). □

Application III (Robustified t-test under Contamination)

Summary: good results (with median and MAD)

- A very nice **pivotal** result which guarantees that **only one** distribution is needed for a given sample size.
- A decent asymptotic result which guarantees the asymptotic normality.

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

- But, speed of convergence is slow as will be shown.

Application III (Robustified t-test under Contamination)

Some questions

- The above is an asymptotic result. Thus, it goes to $N(0,1)$ as $n \to \infty$. In practice, then, how can we use the above result?
- We know that $T_{\mathrm{A}} \xrightarrow{d} N(0,1)$. How quickly does it converge?
- Can we use other distribution as an approximation?

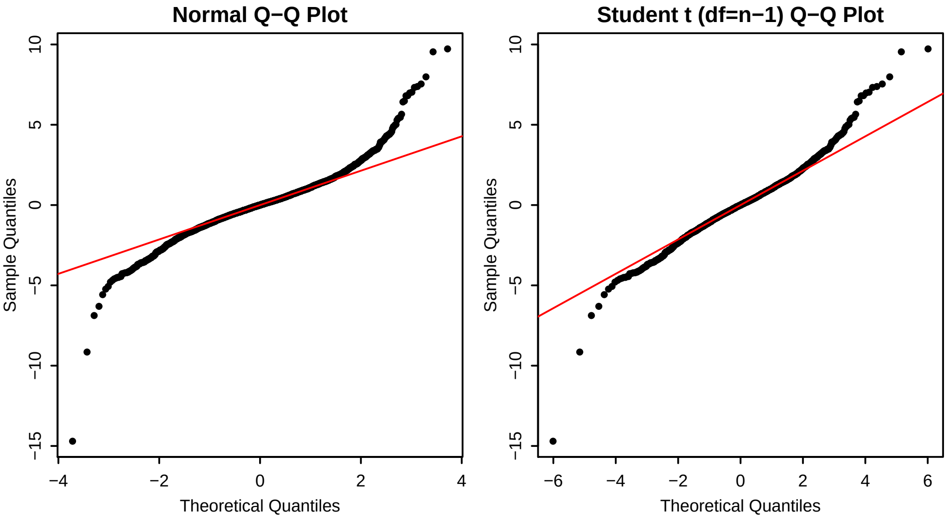


Figure 2: A random sample of size $\boxed{n = 10}$ and iteration 5,000. (a) T_{mM} versus $N(0, 1)$ quantiles. (b) T_{mM} versus Student t (df=9) quantiles.

Application III (Robustified t-test with median and MAD)

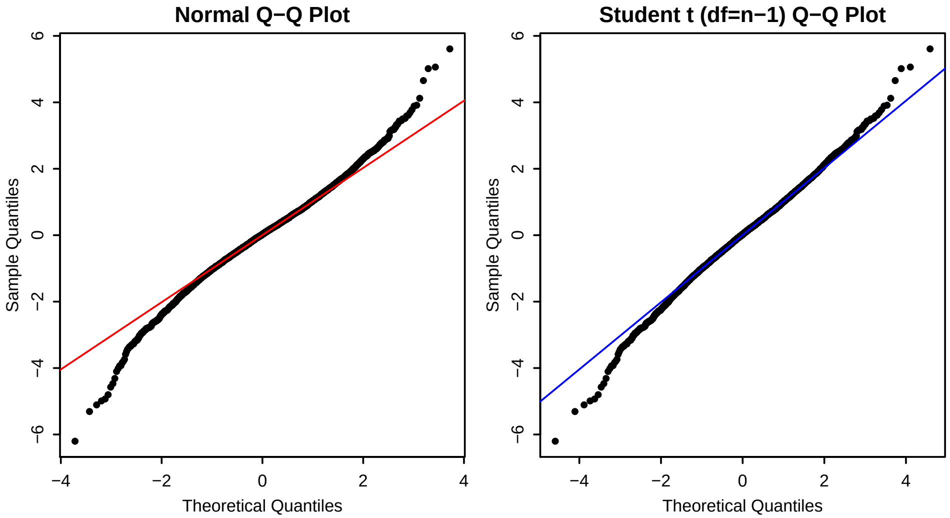


Figure 3: A random sample of size $\boxed{n = 20}$ and iteration 5,000. (a) T_{mM} versus $N(0, 1)$ quantiles. (b) T_{mM} versus Student t (df=19) quantiles.

Application III (Robustified t-test with median and MAD)

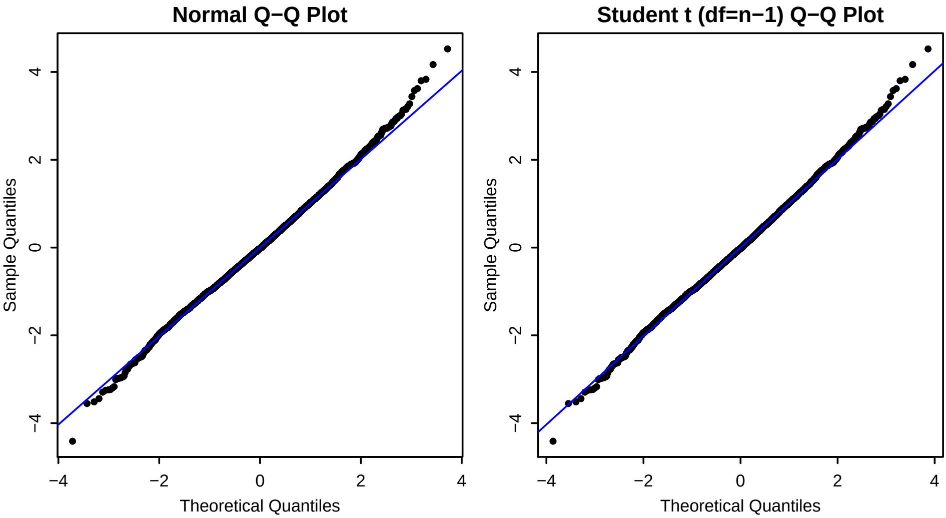


Figure 4: A random sample of size $\boxed{n = 100}$ and iteration 5,000. (a) T_{HS} versus $N(0,1)$ quantiles. (b) T_{HS} versus Student t (df=99) quantiles.

Application III (Hodges-Lehmann and Shamos)

Theorem 3 (Pivot with Hodges-Lehmann and Shamos)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a location-scale family with location μ and scale σ. Then the statistic below is a **pivotal** *quantity:*

$$\frac{\underset{i\le j}{\mathrm{median}}\left(\dfrac{X_i+X_j}{2}\right)-\mu}{\underset{i\le j}{\mathrm{median}}\left(|X_i-X_j|\right)/\sqrt{n}}=\frac{\hat\mu_H-\mu}{\hat\sigma_S/\sqrt{n}}=\frac{\sqrt{n}(\hat\mu_H-\mu)}{\hat\sigma_S}, \tag{6}$$

where $\hat\mu_H$ and $\hat\sigma_S$ are the Hodges-Lehmann and Shamos estimators.

Proof.

See Park (2018) and Jeong et al. (2018). □

Application III (Robustified t-test with under Contamination)

Theorem 4 (Asymptotic Normality with HL and Shamos)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$. Then the following converges to $N(0,1)$.

$$T_B = \sqrt{\frac{6n}{\pi}}\Phi^{-1}(3/4)\frac{\underset{i \le j}{\mathrm{median}}\left(\frac{X_i + X_j}{2}\right) - \mu}{\underset{i \le j}{\mathrm{median}}\left(|X_i - X_j|\right)}$$

$$= \sqrt{\frac{6}{\pi}}\Phi^{-1}(3/4)\frac{\sqrt{n}(\hat{\mu}_H - \mu)}{\hat{\sigma}_S}.$$

Proof.

See Park (2018) and Jeong et al. (2018). □

Application III (Robustified t-test under Contamination)

Summary: good results (with Hodges-Lehmann and Shamos)

- A very nice pivotal result which guarantees that **only one** distribution is needed for a given sample size.
- A very nice asymptotic result which guarantees the asymptotic normality.

$$T_B = \sqrt{\frac{6}{\pi}}\Phi^{-1}(3/4)\frac{\sqrt{n}(\hat{\mu}_H - \mu)}{\hat{\sigma}_S} \xrightarrow{d} N(0,1).$$

Some questions

- The above is an asymptotic result. Thus, it goes to $N(0,1)$ as $n \to \infty$. In practice, then, how can we use the above result?
- We know that $T_B \xrightarrow{d} N(0,1)$. How quickly does it converge?
- Can we use other distribution as an approximation?

Application III (Robustified t-test under Contamination)

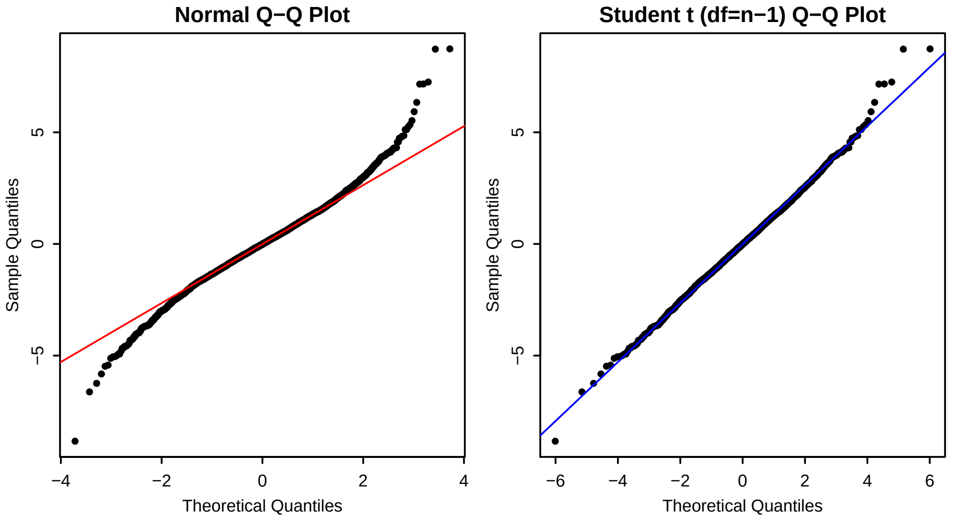


Figure 5: A random sample of size $\boxed{n = 10}$ and iteration 5,000. (a) T_{HS} versus $N(0,1)$ quantiles. (b) T_{HS} versus Student t (df=9) quantiles.

Application III (Robustified t-test under Contamination)

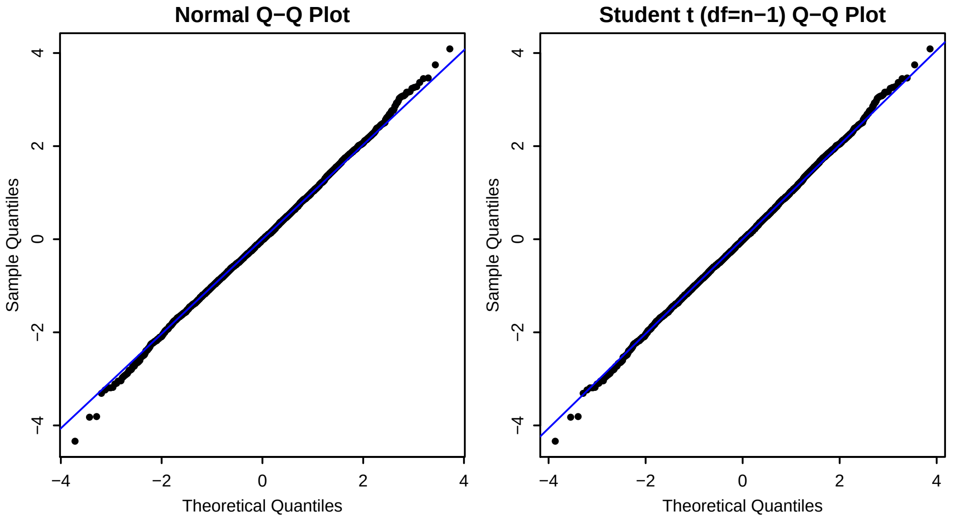


Figure 6: A random sample of size $\boxed{n = 100}$ and iteration 5,000. (a) T_{HS} versus $N(0,1)$ quantiles. (b) T_{HS} versus Student t (df=99) quantiles.

Real Data Example: Darwin's Zea Mays (Corns)

Darwin's Data

- Darwin's Zea Mays (Darwin, 1876) experiment comparing the growth of pairs of corn (especially zea may) seedings, one produced by **self-fertilization** and the other produced by **cross-fertilization**.
- He selected one cross-fertilized plant one self-fertilized plant, grew them in the same pot, and measured their heights.
- This data set has been frequently used by many authors including Fisher (1936), Andrews and Herzberg (1985) among others. See also Section 4.5 of Hogg et al. (2013) and Odiase and Ogbonmwan (2007).
- Let x_i and y_i be the heights of the cross- and self-fertilized plants, respectively, and $d_i = x_i - y_i$
- We want to test

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

Real Data Example: Darwin's Zea Mays (Corns)

Robustified t-test (rt.test) R Package (**empirical distribution**) using Park and Wang (2018a).

https://cran.r-project.org/web/packages/rt.test/

Note that $n = 15 << 100$. (Inappropriate to use asymptotics).

R Exercise with the Darwin's Data

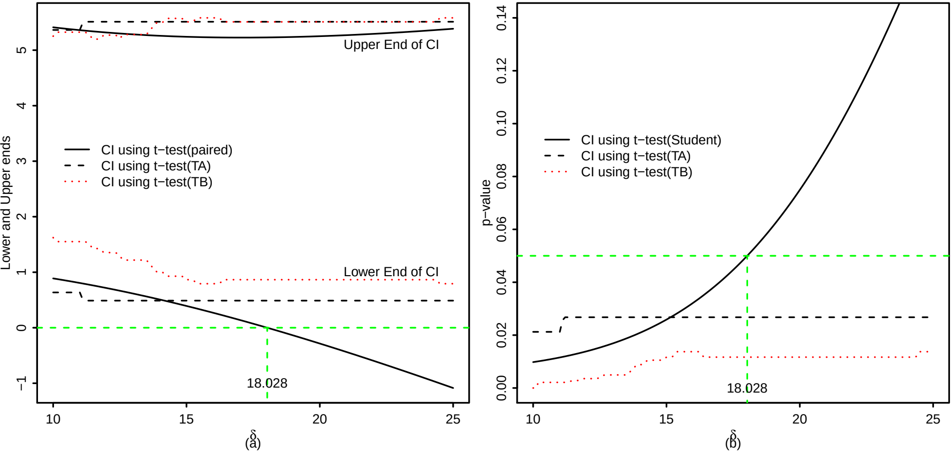
```
install.packages("rt.test")
library(rt.test)
X = c(23.5, 12, 21, 22, 19.125, 21.5, 22.125, 20.375,
      18.25, 21.625, 23.25, 21, 22.125, 23, 12)
Y = c(17.375, 20.375, 20, 20, 18.375, 18.625, 18.625,
      15.25, 16.5, 18, 16.25, 18, 12.75, 15.5, 18)
d = X-Y
 t.test(d)
rt.test(d)
```

- See Talk-2-Example.r on https://github.com/AppliedStat/seminar/tree/master/2018/R

Real Data Example: Darwin's Zea Mays (Corns)

Robustness property of the conventional t-test

- We changed the last value (18) with δ which ranges from 10 to 25 in a grid-like fashion.
- Actually, when the last value (**18**) is replaced with **18.028**, the decision is reversed. That is, the null H_0 becomes accepted from rejection. The difference is only **0.028**
- We also obtained the confidence intervals which say the same story. That is, zero was not included in the interval with the original observation. However, with **0.028** increment, the CI starts to include zero.
- We also calculate the p-values. With **0.028** increment, the p-value starts to increase to 0.05 (5%) or more.
- In what follows, we plot the CIs and p-values of the conventional t-test and proposed robustified t-test.



See Talk-2-Example.r on
`https://github.com/AppliedStat/seminar/tree/master/2018/R`
for the R code.

References

Andrews, D. and Herzberg, A. (1985). *Data: a collection of problems from many fields for the student and research worker*. Springer, New York.

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

Fisher, R. A. (1936). *The design of Experiments*. Oliver and Boyd, London.

Hogg, R. V., McKean, J. W., and Craig, A. T. (2013). *Introduction to Mathematical Statistics*. Pearson, Boston, MA, 7 edition.

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, D. G. and Park, C. (2017). On the generalized mean square error and its applications. Presented at KIIE Conference, Daejon. Nov. 4, 2017.

References

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer, New York.

Odiase, J. I. and Ogbonmwan, S. M. (2007). Exact permutation algorithm for paired observations: The challenge of R.A.Fisher. *Journal of Mathematics and Statistics*, 3:116–121.

Park, C. (2013). Determination of the joint confidence region of optimal operating conditions in robust design by bootstrap technique. *International Journal of Production Research*, 51:4695–4703.

Park, C. (2018). 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.

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

References

Park, C., Ouyang, L., Byun, J.-H., and Leeds, M. (2017). Robust design under normal model departure. *Computers & Industrial Engineering*, 113:206–220.

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

Park, C. and Wang, M. (2018b). `rt.test`: Robustified t-test. `https://CRAN.R-project.org/package=rt.test`. R package version 1.18.7.9.

Serfling, R. J. (2011). Asymptotic relative efficiency in estimation. In Lovric, M., editor, *Encyclopedia of Statistical Science, Part I*, pages 68–82. Springer-Verlag, Berlin.