Parametric versus Nonparametric and Robustness

Chanseok Park (박찬석)

Applied Statistics Laboratory
Department of Industrial Engineering
Pusan National University

August 31, 2018

Hosted by SEC

부산대학교
PUSAN NATIONAL UNIVERSITY

Overview

Overview

Overview

Overview

Overview

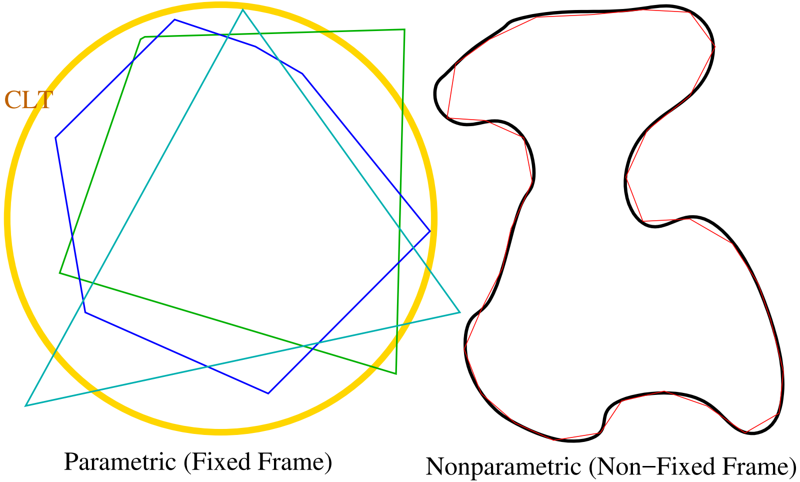
Definitions of Parametric and Nonparametric

Definition

- **Parametric**: Distribution with parameter is assumed.
  Parametric distribution exists.
  ◊ NB: the number of parameters should be fixed.
- **Nonparametric**: Distribution-free.
  No parameters are assumed.
  Parametric distribution does **not** exist.
  - In general, it is acceptable. But, not always.
  - For example, KDE uses the standard normal distribution.

  ◊ NB: the number of parameters depends on the sample size.

Understanding Parametric and Nonparametric concepts



Myths on Parametric and Nonparametric

The myths mentioned here are generally acceptable, but be careful.

- **Nonparametric is robust.**
  Wilcoxon test and KDE are **non-robust** (we will see soon).
  Range ($\max - \min$) is actually **anti-robust**.
- **Mean is parametric and median is nonparametric.**
  - Median is the **parametric MLE** ($\hat{\mu}$) of the Laplace (double-exponential) distribution with $f(x) = \frac{1}{2\sigma}\exp\big(-|x-\mu|/\sigma\big)$.
  - Mean is the MLE of some distributions including the normal.
    But, **without a distribution**, we can view it as a representative value of population. Then, **the mean is nonparametric**.
- **Parametric is non-robust or anti-robust.**
  As seen above, the MLE of the Laplace is **robust**.
  The MLEs of Cauchy and t-dist. are also **robust**.
- **Parametric does NOT work well with non-normal distribution**
  If sample size is decent, it generally **works well** due to CLT.
  That is, robust to model departure (Prof. Byun at GNU).

Myths on Parametric and Nonparametric

- **Nonparametric method does not need any parameters.**
  - KDE has a bandwidth (each kernel is a pdf – usually normal pdf).
    NB: The number of kernels in KDE is the same as the sample size (that is, the number is not fixed).
    NB: If the number of kernels is fixed, it is like a parametric mixture model.
  - Parametric bootstrap method needs parameters.
- **Histogram is nonparametric.**
  - The number of bins, $k \sim \sqrt{n}$ (that is, k is not fixed).
  - However, if the number of bins is fixed (k is fixed), then it is **parametric** multinomial. (Say, $k = 2$, it is binomial).
- **MLE is parametric.**
  - Kaplan-Meier is the MLE but nonparametric.
    (the number of estimates is NOT fixed).
  - The MLE from *empirical likelihood* is nonparametric.
    (the number of estimates is NOT fixed).

Parametric versus Nonparametric

Comparison: parametric versus Nonparametric

Description	Parametric	Nonparametric	Note
Popular estimate	mean	median	in general
Robustness	bad	good	in general
Power of test	good	bad	
Model departure	not bad	good	NB:CLT
Sample size	needs big size	small is OK	
Calculation	easier	more complex	
# of parameters	fixed	depends on sample size	
Information loss	minor	can be serious	NB:rank,median
Examples	MLE, MME, Bayesian pdf with parameter	Range, kernel method, empirical likelihood, bootstrap, plots (hist, etc)	
Others		bad with tied or zero values	

Robustness Issues (Example: Mean versus Median)

In general, mean is regarded as a parametric estimate and median as a nonparametric one. Thus, we will investigate mean and median in details.

Example (Refer to Talk-2 at github.com/AppliedStat/seminar)

- Data: $Y = (-2, -1, 0, 1, 2)$: mean $= 0$ and median $= 0$.
- Data: $Y = (-2, -1, 0, 1, 102)$: mean $= 20$ and **median $= \mathbf{0}$**.

- The **mean** loves all including an evil.
  - If there is **no** evil, the mean is usually better.
  - Because it loves all, it can be easily fooled by a sweet evil.
  - NB: Mean has zero breakdown (Refer to Talk-2).
- The **median** loves the middle only.
  - It ignores all the values except the middle one.
  - If there is an evil, the median is usually better. (high breakdown point).
  - It loses valuable information (efficiency issue) because it ignores all the observations except the middle one.

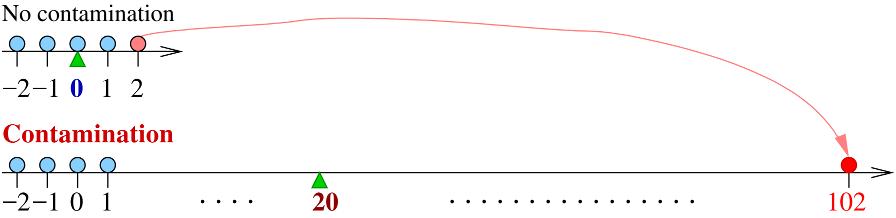
◇ NB: Two important measures: **breakdown** and **relative** efficiency.

Refer to Talk-2 and Talk-6 at github.com/AppliedStat/seminar.

Robustness Issues (Geometry on Mean and Median)

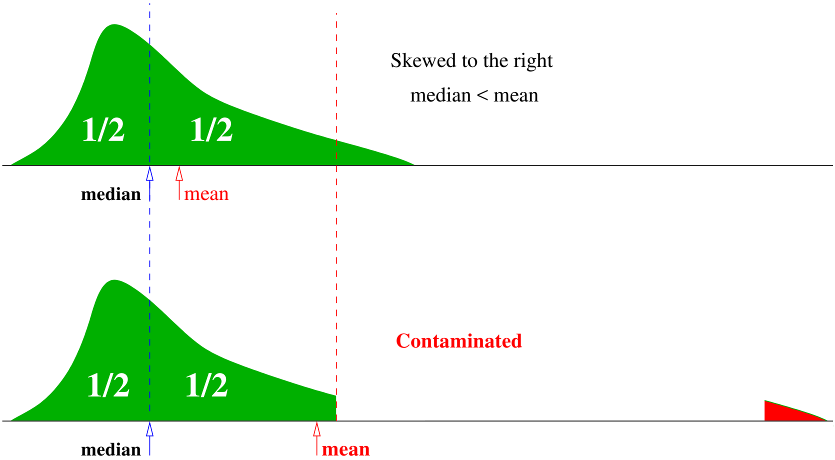
Why the mean is not robust? Recall mean: $\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$

- Data: $Y = (-2, -1, 0, 1, 2)$: mean $= 0$ and median $= 0$
- Data: $Y = (-2, -1, 0, 1, 102)$: mean $= 20$ and **median $= 0$**



The mean is the center of gravity while the median is just the middle one.
The mean is influenced by the gravity (leverage) while the median is NOT.

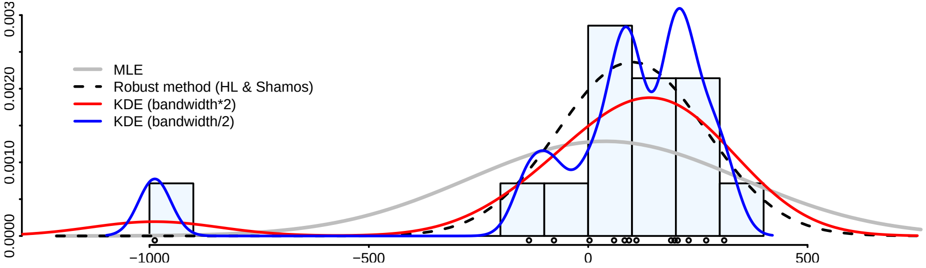
Robustness Issues (Geometry on Mean and Median)



- The mean is the center of **gravity** of pdf ~~pizza~~.
- The median is the center of **area** (half-half area) of pdf ~~pizza~~.

Robustness Issues (PDF and KDE)

The difference between the two faults rate (test and control phone-lines) from Welch (1987).



- −988, −135, −78, 3, 59, 83, 93, 110, 189, 197, 204, 229, 269, 310.
- The KDEs are not robust to the outlier although it is not so lethal. The bump shape changes with the bandwidth, but the area around the outlier is not deleted. Also, we have a bandwidth selection problem.
- The pdf with the MLE will collapse as the outlier (−988) moves to the left.
- The pdf with the robust estimates will keep the same shape even with the change of the outlier (−988).

Example: Paired two-sample t-test

Biology data for self- versus cross-fertilization by Darwin (1876)

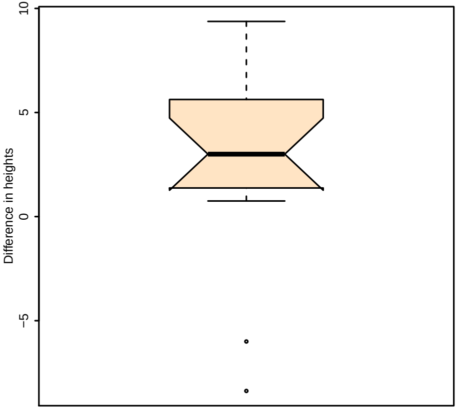
Cross	23.500	12.000	21	22	19.125	21.500	22.125	20.375
	18.25	21.625	23.25	21	22.125	23.0	12	
Self	17.375	20.375	20	20	18.375	18.625	18.625	15.250
	16.50	18.000	16.25	18	12.750	15.5	**18**	

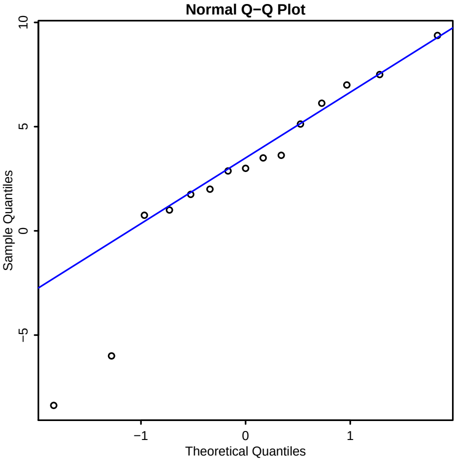
- Let $D_i = X_i - Y_i$ where X_i and Y_i are the heights of the cross- and self-fertilized plants.
- Hypothesis test: $H_0 : \mu_d = 0$ versus $H_1 : \mu_d \neq 0$, where $\mu_d = \mu_X - \mu_Y$.
- We will compare (i) paired t-test, (ii) nonparametric Wilcoxon test (Wilcoxon, 1945), (iii) Yuen's t-tes (Yuen and Dixon, 1973), and (iv) rt-test (Park and Wang, 2018a,b).
- The last value ($Y_{15} = \mathbf{18}$) will be contaminated.

◇ This Example will be presented at KIIE, Seoul.

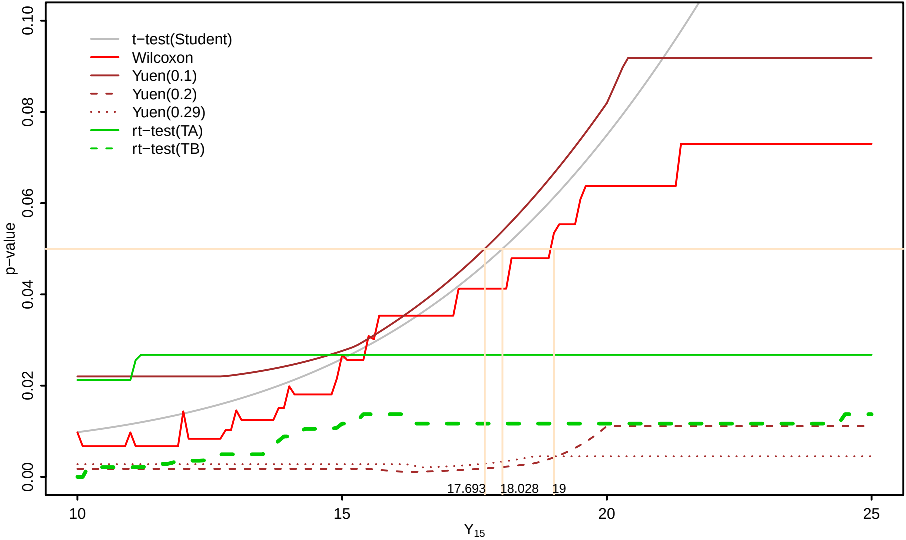
◇ Also, see $\boxed{\text{Talk-2}}$ (confidence interval) at seminar/R/Talk-2.r.

Example: Paired two-sample t-test





Example: Paired two-sample t-test



Warnings from in `wilcox.test()` R function

Note: The R codes for the plots in this Talk are posted at ▸ seminar/R/Talk-4.r

Warning messages from `wilcox.test()` R function

```
Warning messages:
1: In wilcox.test.default(X, Ynoised, paired = TRUE) :
  cannot compute exact p-value with ties
2: In wilcox.test.default(X, Ynoised, paired = TRUE) :
  cannot compute exact p-value with ties
3: In wilcox.test.default(X, Ynoised, paired = TRUE) :
  cannot compute exact p-value with zeroes
4: In wilcox.test.default(X, Ynoised, paired = TRUE) :
  cannot compute exact p-value with ties
5: In wilcox.test.default(X, Ynoised, paired = TRUE) :
  cannot compute exact p-value with ties
 ........
```

As mentioned earlier, Wilcoxon test (nonparametric) has a problem with dealing with tied or zero values.

References

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

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.

Welch, W. J. (1987). Rerandomizing the median in matched-pairs designs. Biometrika, 74:609–614.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1:80–83.

Yuen, K. K. and Dixon, W. J. (1973). The approximate behaviour and performance of the two-sample trimmed t. Biometrika, 60:369–374.