Quality Control with Unequal Sample Sizes (Other Applications)

Chanseok PARK[1] Linhan OUYANG[2] Min WANG[3]

[1]Applied Statistics Laboratory
Department of Industrial Engineering, Pusan National University

[2]College of Economics and Management
Nanjing University of Aeronautics and Astronautics

[3]Department of Management Science and Statistics
The University of Texas at San Antonio

December 3, 2020

Hosted by NUAA

Overview

Overview

Overview

Overview

Recall

Five methods

- $\bar{S}_A = \dfrac{1}{m}\sum_{i=1}^{m}\dfrac{S_i}{c_4(n_i)}$
- $\bar{S}_B = \dfrac{\sum_{i=1}^{m} S_i}{\sum_{i=1}^{m} c_4(n_i)}$
- $\bar{S}_C = \dfrac{\sum_{i=1}^{m}\dfrac{c_4(n_i)}{1-c_4(n_i)^2}\cdot S_i}{\sum_{i=1}^{m}\dfrac{c_4(n_i)^2}{1-c_4(n_i)^2}}$ $\Leftarrow$ **BEST**
- $\bar{S}_D = S_p/c_4(N-m+1)$ Note: less power under heteroscedasticity.
- $\bar{S}_E = S_N/c_4(N)$ Note: less power under heteroscedasticity or H_1.

BLUE (best linear unbiased estimator)

Theorem 1 (BLUE)

The estimator $\bar{S}_C$ is the BLUE.

Sketch proof.

Consider a linear unbiased estimator in the form of $\sum_{i=1}^{m} w_i S_i$. The variance and expectation are given by

$$\mathrm{Var}\Big(\sum_{i=1}^{m} w_i S_i\Big) = \sum_{i=1}^{m} w_i^2 \{1 - c_4(n_i)^2\}\sigma^2$$

$$E\Big(\sum_{i=1}^{m} w_i S_i\Big) = \sum_{i=1}^{m} w_i c_4(n_i)\sigma.$$

We need minimize $\mathrm{Var}\left(\sum_{i=1}^{m} w_i S_i\right)$ with the unbiasedness condition $E\left(\sum_{i=1}^{m} w_i S_i\right) = \sigma$.

BLUE (best linear unbiased estimator)

Thus, our objective is to minimize

$$\sum_{i=1}^{m} w_i^2\{1 - c_4(n_i)^2\} \quad \text{subject to} \quad \sum_{i=1}^{m} w_i c_4(n_i) = 1,$$

which can be easily solved by using the method of Lagrange multipliers. □

Can we have more general results for BLUE?

Dr. Ouyang, Dr. Wang and I are currently working on this issue.

We briefly introduce these results.
First we will review $\bar{\bar{X}}_A$ and $\bar{\bar{X}}_B$ and then provide the results.

BLUE (best linear unbiased estimator)

Recall $\bar{\bar{X}}_A$ and $\bar{\bar{X}}_B$

$$\bar{\bar{X}}_A = \frac{\bar{X}_1 + \bar{X}_2 + \cdots + \bar{X}_m}{m} = \frac{1}{m}\sum_{i=1}^{m} \bar{X}_i \tag{1}$$

and

$$\bar{\bar{X}}_B = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \cdots + n_m\bar{X}_m}{n_1 + n_2 + \ldots + n_m} = \frac{1}{N}\sum_{i=1}^{m} n_i\bar{X}_i, \tag{2}$$

where $\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$ and $N = \sum_{i=1}^{m} n_i$.
Using the inequalities of the AM and HM, Park and Wang (2020a) proved

$\mathrm{Var}(\bar{\bar{X}}_A) \geq \mathrm{Var}(\bar{\bar{X}}_B)$.

Like $\bar{\bar{X}}_A$ and $\bar{\bar{X}}_B$, does this inequality work for others (median, say)?

BLUE (best linear unbiased estimator)

One can also estimate the population mean using location estimators as follows

$$\bar{\hat{\mu}}_A = \frac{\hat{\mu}_1 + \hat{\mu}_2 + \cdots + \hat{\mu}_m}{m} = \frac{1}{m}\sum_{i=1}^{m} \hat{\mu}_i \tag{3}$$

and

$$\bar{\hat{\mu}}_B = \frac{n_1\hat{\mu}_1 + n_2\hat{\mu}_2 + \cdots + n_m\hat{\mu}_m}{n_1 + n_2 + \ldots + n_m} = \frac{1}{N}\sum_{i=1}^{m} n_i\hat{\mu}_i, \tag{4}$$

where $\hat{\mu}_i$ denotes the unbiased location estimator of μ with the ith sample.

It is easily seen that $\bar{\hat{\mu}}_A$ and $\bar{\hat{\mu}}_B$ are unbiased. But, $\mathrm{Var}(\bar{\hat{\mu}}_A) \geq \mathrm{Var}(\bar{\hat{\mu}}_B)$ **does not** hold for some cases (median, say). $\mathrm{Var}(\bar{\bar{X}}_A) \geq \mathrm{Var}(\bar{\bar{X}}_B)$ is a special case. We will look at some general results.

BLUE (best linear unbiased estimator)

Theorem 2 (BLUE for general location and scale estimators)

The BLUE for the location is given by

$$\bar{\hat{\mu}} = \frac{\sum_{i=1}^{m}(\hat{\mu}_i/\nu_i^2)}{\sum_{i=1}^{m}(1/\nu_i^2)}, \tag{5}$$

where ν_i^2 is the variance of $\hat{\mu}_i$ under the normal distribution.
The BLUE for the scale is given by

$$\bar{\hat{\sigma}} = \frac{\sum_{i=1}^{m}(\gamma_i/\tau_i^2)\hat{\sigma}_i}{\sum_{i=1}^{m}(\gamma_i^2/\tau_i^2)}, \tag{6}$$

where γ_i and τ_i^2 are the expectation and variance of $\hat{\sigma}_i$ under the normal distribution, respectively.

We can prove the above using the method of Lagrange multipliers.

BLUE (best linear unbiased estimator)

As an example of the location estimator, let's apply Theorem 2 for the mean estimator. We have $\mathrm{Var}(\bar{X}_i) = \sigma_i^2/n_i$.

$$\bar{\hat{\mu}} = \frac{\sum_{i=1}^m (\bar{X}_i/(\sigma_i^2/n_i)}{\sum_{i=1}^m 1/(\sigma_i^2/n_i)}.$$

With the assumption that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2$, we have

$$\bar{\hat{\mu}} = \frac{\sum_{i=1}^m n_i \bar{X}_i}{\sum_{i=1}^m n_i} = \frac{1}{N}\sum_{i=1}^m n_i \bar{X}_i = \bar{\bar{X}}_B.$$

For the mean, $\bar{\bar{X}}_B$ is the BLUE.

In general, however, $\bar{\hat{\mu}}_B = \frac{1}{N}\sum_{i=1}^m n_i \hat{\mu}_i$ is NOT BLUE.

BLUE (best linear unbiased estimator)

Under $X_{ij} \sim N(\mu, \sigma^2)$ where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n_i$, $\mathrm{Var}(\bar{\bar{X}}_A) \geq \mathrm{Var}(\bar{\bar{X}}_B)$ always. However the median is used, the inequality does not work.

For example, let $\hat{\mu}_i$ be the median of the ith sample with size n_i.

With $n_1 = 4$ and $n_2 = 5$, we have

$$\mathrm{Var}(\bar{\hat{\mu}}_A) < \mathrm{Var}(\bar{\hat{\mu}}_B)$$

Recall:

$$\bar{\hat{\mu}}_A = \frac{\hat{\mu}_1 + \hat{\mu}_2 + \cdots + \hat{\mu}_m}{m} = \frac{1}{m} \sum_{i=1}^{m} \hat{\mu}_i$$

and

$$\bar{\hat{\mu}}_B = \frac{n_1\hat{\mu}_1 + n_2\hat{\mu}_2 + \cdots + n_m\hat{\mu}_m}{n_1 + n_2 + \ldots + n_m} = \frac{1}{N} \sum_{i=1}^{m} n_i\hat{\mu}_i.$$

BLUE (best linear unbiased estimator)

- What is the problem with the above $\bar{\hat{\mu}}_A$ and $\bar{\hat{\mu}}_B$?
- It is well known that the variance of the sample mean and median are $\mathrm{Var}(\bar{X}_i) = \sigma^2/n_i \propto 1/n_i$ and $\mathrm{Var}(\hat{\mu}_i) \approx (\pi/2) \cdot (\sigma^2/n_i) \propto 1/n_i$ (for large n).
- Is Theorem 2 OK? Yes, Theorem 2 is OK.
- Note that $\mathrm{Var}(\bar{X}_i) \propto 1/n_i$ but $\mathrm{Var}(\hat{\mu}_i) \not\propto 1/n_i$ (for small n).
- What should we do for small n.
  For more details on $\mathrm{Var}(\hat{\mu}_i)$ (with small sample), refer to Park et al. (2021) and Park and Wang (2020b) which calculated the exact variance of the sample median with small sample.
- Why is the median needed? It has a positive breakdown point.
- $\mathrm{Var}(\bar{\bar{X}}_A) \geq \mathrm{Var}(\bar{\bar{X}}_B)$ works because $\mathrm{Var}(\bar{X}_i) \propto 1/n_i$.

mean/SD versus median/IQR

Sample mean and variance

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ (mean)} \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \text{ with } S = \sqrt{S^2} \text{ (SD)}.$$

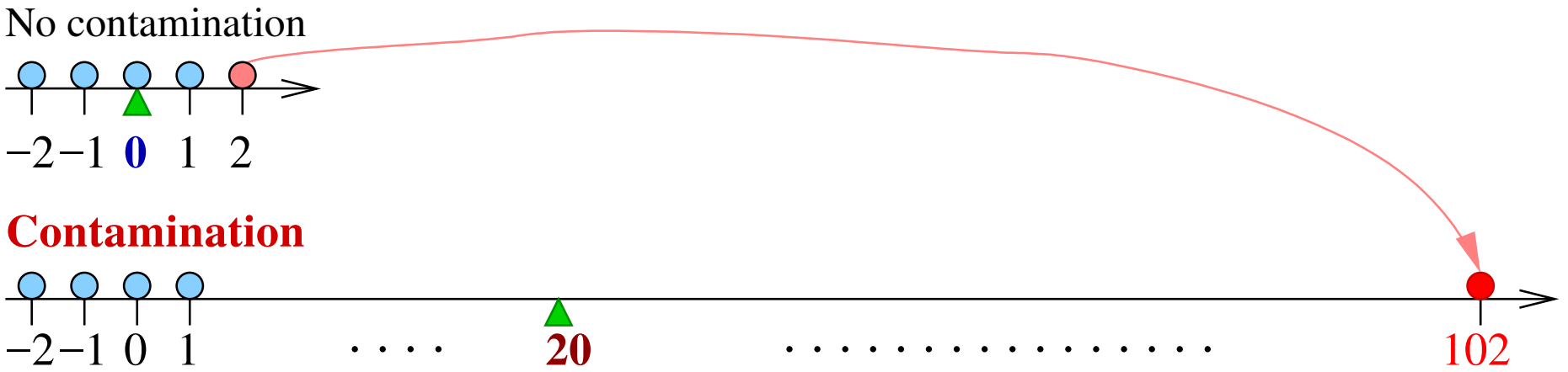
Illustrative Example

	Original data $(-2, -1, 0, 1, 2)$	Contaminated data $(-2, -1, 0, 1, 102)$
Mean	0	**20**
Median	0	0
SD	1.58	**45.9**
IQR	2	2

View from physics (mean vs. 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.

BLUE with scale estimator

Using Theorem 2, we can make a pooled scale estimator. We briefly review the pooled sample variance. We have

$$\frac{(n_i-1)S_i^2}{\sigma^2} \sim \chi^2(\mathrm{df}=n_i-1) \quad \text{under the normality.}$$

We have

$$E\left[\frac{(n_i-1)S_i^2}{\sigma^2}\right] = n_i - 1 \quad \text{and} \quad \mathrm{Var}\left[\frac{(n_i-1)S_i^2}{\sigma^2}\right] = 2(n_i-1).$$

Thus, we have

$$\boxed{\mathrm{Var}(S_i^2) = \frac{2\sigma^4}{n_i-1} \propto \frac{1}{n_i-1}.}$$

Using the above and substituting it into (6), we have

$$S_p^2 = \frac{(n_1-1)S_1^2 + \cdots + (n_m-1)S_m^2}{n_1+\cdots+n_m-m}$$

BLUE with scale estimator

- S_p^2 is BLUE for σ^2 and unbiased for σ^2.
- But, $S_p = \sqrt{S_p^2}$ is NOT BLUE. It is also biased.
- $$\bar{S}_C = \frac{\sum_{i=1}^{m} \frac{c_4(n_i)}{1 - c_4(n_i)^2} \cdot S_i}{\sum_{i=1}^{m} \frac{c_4(n_i)^2}{1 - c_4(n_i)^2}} \quad \Leftarrow \textbf{BLUE} \text{ and unbiased}$$

  Actually, $\bar{S}_C$ in the above is derived using Theorem 2.

BLUE

Pooling is very important, which can create pooled estimators including the **grand mean** and **pooled sample variance**.

- To estimate location parameter (grand mean)

$$\boxed{\bar{\bar{X}}_B = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \cdots + n_m\bar{X}_m}{n_1 + n_2 + \ldots + n_m} = \frac{1}{N}\sum_{i=1}^{m} n_i\bar{X}_i}$$

  is preferred over $\bar{\bar{X}}_A = (\bar{X}_1 + \bar{X}_2 + \cdots + \bar{X}_m)/m$

- To estimate

$$\boxed{S_p^2 = \frac{(n_1 - 1)S_1^2 + \cdots + (n_m - 1)S_m^2}{n_1 + \cdots + n_m - m}}$$

  is preferred (BLUE).

- To estimate σ, however, $\bar{S}_C = \sum_{i=1}^{m} \frac{c_4(n_i)}{1-c_4(n_i)^2} \cdot S_i \Big/ \sum_{i=1}^{m} \frac{c_4(n_i)^2}{1-c_4(n_i)^2}$ is preferred (BLUE), not S_p.

BLUE for other estimators

- The grand mean and pooled sample variance are widely used, but they are NOT robust at all.
- Thus, we need to use robust estimators to make pooled estimators.
- But, special cares should be taken to ensure the **unbiasedness** of estimators and **best performance**.
- Let $\hat{\mu}_i$ be the median. The below are NOT appropriate.

$$\bar{\hat{\mu}}_A = \frac{\hat{\mu}_1 + \hat{\mu}_2 + \cdots + \hat{\mu}_m}{m} \quad \text{and} \quad \bar{\hat{\mu}}_B = \frac{n_1\hat{\mu}_1 + n_2\hat{\mu}_2 + \cdots + n_m\hat{\mu}_m}{n_1 + n_2 + \ldots + n_m}$$

  $\Rightarrow$ Use Theorem 2 along with Park et al. (2021) (this is needed for proper weights).

  $\Rightarrow$ Dr. Ouyang, Dr. Wang and I are working on this .

- Let $\hat{\sigma}_i$ be the MAD (robust).
  Then how to pool? The below is NOT appropriate.

$$\bar{\hat{\sigma}}_B = \frac{(n_1 - 1)\hat{\sigma}_1 + (n_2 - 1)\hat{\sigma}_2 + \cdots + (n_m - 1)\hat{\sigma}_m}{n_1 + n_2 + \ldots + n_m - m}$$

Applications

Robust design with unbalanced data (Cho and Park, 2005)

- Let $\hat{m}(\mathrm{x})$ and $\hat{v}(\mathrm{x})$ represent the fitted response functions for the mean and variance of the response Y, respectively.
- Assuming a second-order polynomial model for the response functions, we get

$$\hat{m}(\mathrm{x}) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i + \sum_{i=1}^{k} \sum_{j=i}^{k} \hat{\beta}_{ij} x_i x_j$$

$$\hat{v}(\mathrm{x}) = \hat{\gamma}_0 + \sum_{i=1}^{k} \hat{\gamma}_i x_i + \sum_{i=1}^{k} \sum_{j=i}^{k} \hat{\gamma}_{ij} x_i x_j.$$

- We use the sample mean and variance of Y to estimate the process mean $\hat{m}(\mathrm{x})$ and variance $\hat{v}(\mathrm{x})$, respectively.

Applications

Table 1: Data for case study example.

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

- $\overline{Y}_i$ is used to obtain $\hat{m}(\mathrm{x})$ with **WLS** regression with weights $1/n_i$.
  If the median is used instead of $\overline{Y}_i$, what are appropriate weights?
- S_i^2 is used to obtain $\hat{v}(\mathrm{x})$ with **WLS** with weights $1/(n_i - 1)$.
  If the MAD is used instead of S_i^2, what are appropriate weights?

Summary

Summary

- General Theorem for BLUE is provided for pooling.
- When the sample mean or variance are pooled, it is easy to pool them.
- But, when other estimators are pooled, special cares should be taken for proper weights.
  To this end, the methods in Park et al. (2021) can be considered.
- For various applications including robust design, robust quality control, etc., proper weights are needed.

References

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

Park, C., Kim, H., and Wang, M. (2021). Investigation of finite-sample properties of robust location and scale estimators. Communication in Statistics – Simulation and Computation, To appear. doi:10.1080/03610918.2019.1699114.

Park, C. and Wang, M. (2020a). A study on the X-bar and S control charts with unequal sample sizes. Mathematics, 8(5):698.

Park, C. and Wang, M. (2020b). `rQCC`: Robust quality control chart. `https://CRAN.R-project.org/package=rQCC`. R package version 1.20.7 (published on July 5, 2020).