Frequently Asked Questions

Chanseok Park (박찬석)

Applied Statistics Laboratory
Department of Industrial Engineering
Pusan National University

August 31, 2018

Hosted by SEC

Overview

Overview

Overview

Overview

Overview

1. 비모수 통계에서 confidence interval 도출 방식
2. 신뢰성 data (가속 수명 data) 분석
3. 신뢰도/고장률 예측(light tail)
4. Heavy tail에서 percentile 추정
5. KDE 방법에서 Bandwidth의 선정과 kernel의 선택
6. Robust statistics with high-dimensional data
7. CI with a small sample size

Overview

1. 비모수 통계에서 confidence interval 도출 방식
2. 신뢰성 data (가속 수명 data) 분석
3. 신뢰도/고장률 예측(light tail)
4. Heavy tail에서 percentile 추정
5. KDE 방법에서 Bandwidth의 선정과 kernel의 선택
6. Robust statistics with high-dimensional data
7. CI with a small sample size

Overview

1. 비모수 통계에서 confidence interval 도출 방식
2. 신뢰성 data (가속 수명 data) 분석
3. 신뢰도/고장률 예측(light tail)
4. Heavy tail에서 percentile 추정
5. KDE 방법에서 Bandwidth의 선정과 kernel의 선택
6. Robust statistics with high-dimensional data
7. CI with a small sample size

비모수 통계에서 confidence interval – 1/3

If we have order statistics, $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, then the confidence interval (equitail CI) can be obtained as

$$[x_{(k)}, x_{(m)}].$$

For a very **rough illustration**, to obtain 90% CI with a sample of $x_{(1)}, x_{(2)}, \ldots, x_{(100)}$, we have find the 5th and 10th percentiles (0.05 and 0.95 quantiles).

$$[x_{(5)}, x_{(95)}].$$

- There is a minor adjustment, so-called, plotting position.
- The R function

```
quantile(x, probs=c(0.05, 0.95))
```

will calculate 90% CI with **adjustment**.

비모수 통계에서 confidence interval – 2/3

What if we have only one estimate, say, $\hat{\theta}$?

(very rough idea)

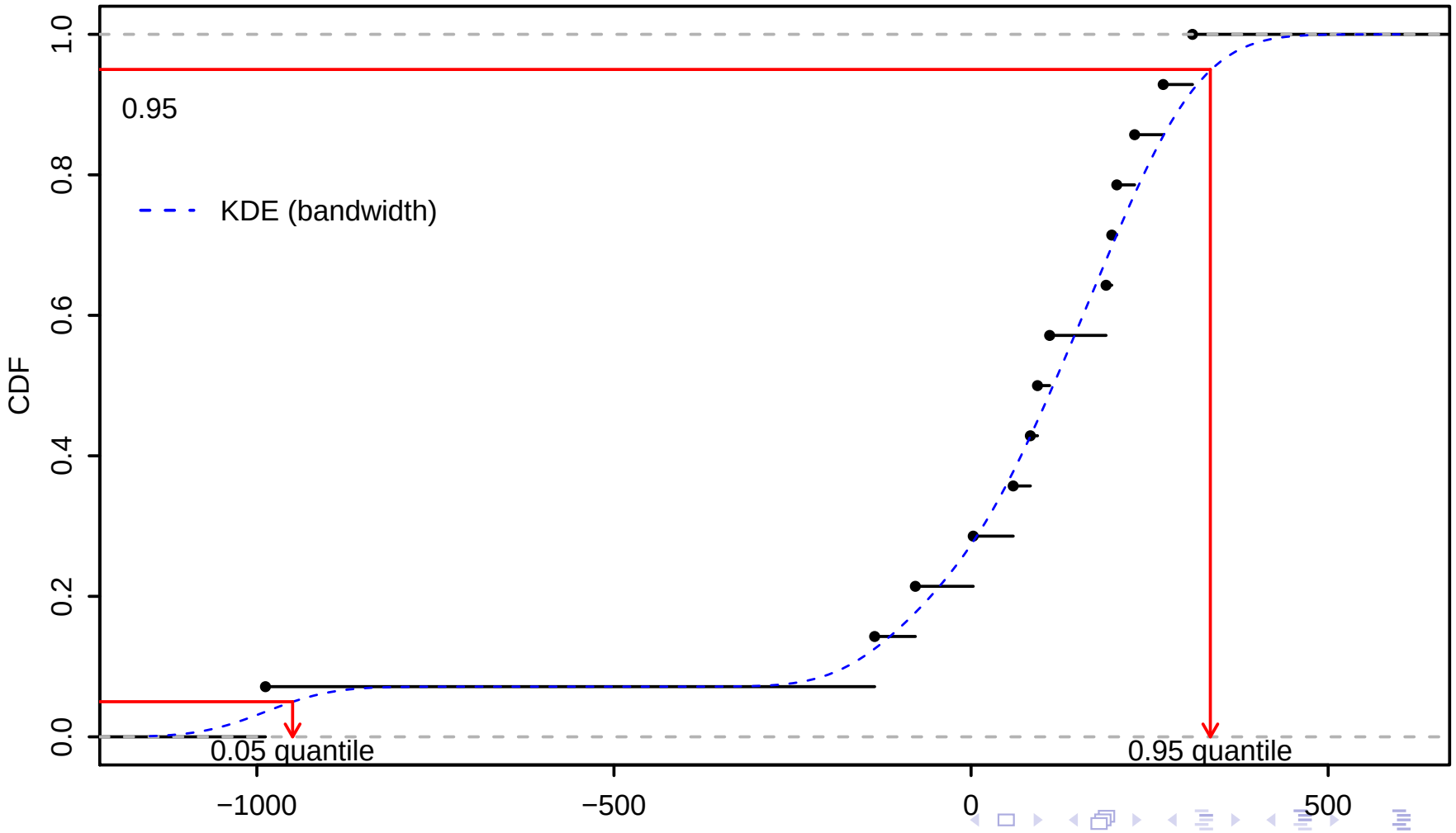
- Then, resample the original sample and then estimate. Let's call it $\hat{\theta}_1$.
- Repeat again and again (say, up to B times). This is called bootstrap resampling.
- Then we have $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B$.
- Sort the above. Then obtain

$$[\hat{\theta}_{(k)}, \hat{\theta}_{(m)}].$$

For a better way, see §2.4 of Davison and Hinkley (1997).

비모수 통계에서 confidence interval – 3/3

What is quantile? It is an inverse cdf.
What if sample size is small? We can use a CDF based on kernel (smooth value).



신뢰성 data (가속 수명 data) 분석 – 1/2

Question: 샘플 수가 작은(100개 이하) 경우, 또는 **censoring** 분석을 하는 경우, 보다 robust한 percentile(0.1%, 0.01% 등) 수명을 추정하는 방법.

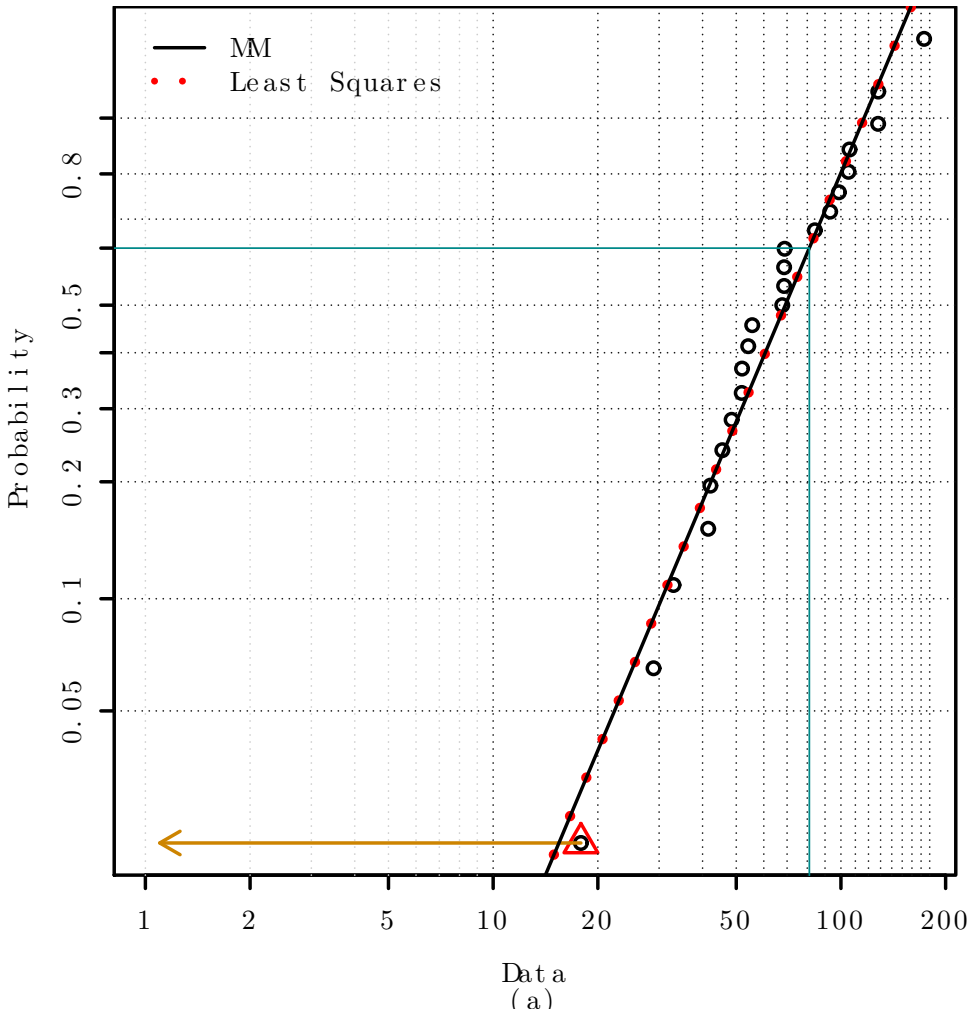
It is very likely that an outlier is in an interval (right censoring). Thus, handling censoring with outlier is really tricky.
Also, we can estimate the parameters from the **Weibull plot**.

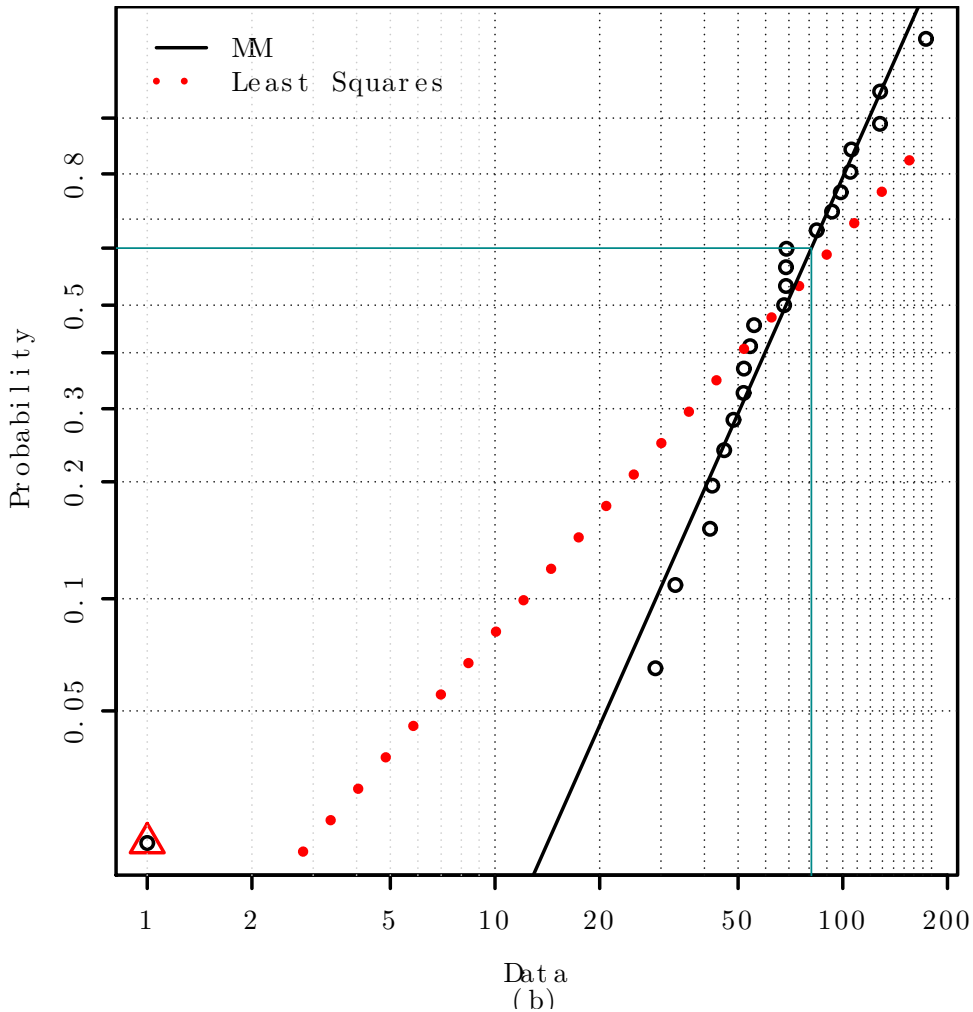
- Estimate an cdf using Kaplan-Meier if there are **censored data**.
- Find the intercept and slope using a robust regression (say, `rlm` R function in MASS package).
- From the slope and intercept, reparametrize and then obtain the Weibull parameter.

What if the distribution is not Weibull?
Use the MLE of a heavy-tailed distribution (say, Laplace).
$\Diamond$NB: the `robust` median is the MLE of the Laplace distribution and the Hodges-Lehmann estimator is the MLE of the logistic distribution (Serfling, 2011).

신뢰성 data (가속 수명 data) 분석 – 2/2

Weibull example:





신뢰도/고장률 예측(light tail)

Question: 신뢰도/고장률 예측 등의 분야에서는 중심치나 산포 보다는 tail에 대한 정밀한 추정이 요구됩니다. heavy tail이 아니라 light tail일 때, 그리고 해당 data는 **contaminated data가 아니라** real data일 때 tail에 위치한 이러한 data들을 보다 잘 반영하는 추정법 등이 궁금합니다.

- It is basically about GOF (goodness of fit) problem. Refer to Talk-3 at ▸ Seminar/2018.
- It is also related with a rare-event distribution.
- Extreme value theory (this also deals with extreme or rare events).
- Developed by earthquake, weather, etc. For example, Gumbel (1963).

Heavy tail에서 percentile 추정

Question: Heavy tail을 갖는 분포(user 수집 data로 샘플 10만 이상)에서 percentile(1%, 0.1%, 0.01% 등)값 추정 방법

(a) 정규 분포로 modeling 하여 modeling 된 정규 분포 기준으로 percentile 값을 추정하는 방법.

(b) 수집 data(샘플 수가 많기 때문에) 기준으로 percentile 값을 직접 확인하는 방법

(a1) It is also basically about GOF (goodness of fit) problem.
For the normality, plot QQ-plot or perform Shapiro-Wilk test.
For example, use `shapiro.test()` R function.
If it is normal, it is OK to use the normal.

(a2) One can use some transforms (say, Box–Cox Transform) to make it normal. Then use existing well-developed methods.

(b) If still not normal? One can use the inverse of an empirical CDF.
See Page 5. One can also superimpose the fitted normal cdf with an empirical cdf to compare the fits.

KDE 방법에서 Bandwidth의 선정과 kernel의 선택 – 1/2

Question:

(a) KDE 방법에서 Bandwidth의 선정이 중요한데 추천하시는 방법이 있으신지요?

(b) Kernel의 선택이 가장 중요한 것 같은데, 효과적인 Kernel 선택방법이 있는지?

(c) 특정 Kernel의 경우 추정해야하는 Parameter 수가 많은데 이런 경우 효과적인 추정방법이 있는지?

(a1) The most well-known method is to select **global** h which minimizes MISE (mean integrated square error)

$$\mathrm{MISE}(\hat{f}) = E\left[\int \left\{\hat{f}(t) - f(t)\right\}^2 dt\right].$$

$\Diamond$NB: $\hat{f}(t)$ is a function of h and a random sample, $\{X_1, X_2, \ldots, X_n\}$. The most famous rule is $\hat{h} \sim \hat{\sigma} n^{-0.2}$. To estimate $\hat{\sigma}$, a robust method is recommended. Refer to Chapter 3 of Silverman (1986).

KDE 방법에서 Bandwidth의 선정과 kernel의 선택 – 2/2

(a2) R function `bw.nrd0(x)`, `bw.nrd(x)`, etc. can be used.
The above methods are based on symmetric kernels.
As far as I know, this selection problem is well settled.
But, for KDE with **asymmetric** kernels (to avoid spill-over effect due to bounded domain, etc.), one can use Cross validation, bootstrap bandwidth selection procedures, etc.

(b) Personally, I prefer to use Epanechnikov kernel because it is defined in a bounded range (between $\pm\sqrt{5}$) and higher efficiency.
See Table 3.1 of Silverman (1986).
As mentioned in (a2) above, if asymmetric kernels are used, it is a very difficult problem.

(c) My understanding is that this is about a local bandwidth selection problem. For a local bandwidth section, h_i is used instead of a global h, (also, along with different weights w_i instead of $1/n$), I am not an expert. Sorry.

Robust statistics with high-dimensional data – 1/3

Question: 로버스트한 통계적 방법론을 사용하는 가장 큰 이유가 이상치(outlier)로 이해하고 있습니다.

(a) 보통 1-dim에서의 데이터들은 boxplot method등으로 IQR 기반으로 제거하는 방법이 일반적인데,

(b) multi, high-dimension 에서의 데이터는 어떻게 outlier를 처리하나요? (KNN distance 등 응용?)

(a1) Personally, IQR is not recommended. See Talk-2 at ▸ Seminar/2018. In robustness, two important measures: ARE and breakdown.
It is not recommended to remove observations manually because they can be an observation from model departure or a surprising observation. Refer to Page 9 of Talk-6.
Let's recall ARE and breakdown briefly.

Robust statistics with high-dimensional data – 2/3

(a2) ARE and breakdown.

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%**

Robust statistics with high-dimensional data – 1/3

(b) Tukey (1975) proposed a multivariate median based on the depth function of a data set.
The depth of a data point is reversely related to its outlyingness.

Möttönen and Oja (1995) also proposed a multivariate median based on rank method.
See also Chapter 6 of Hettmansperger and McKean (2010).

CI with a small sample size

Question: 샘플사이즈가 충분하지 않을 때 confidence interval을 구하기 위하여 resampling을 토대로 한 bootstrap 방법이 사용된다고 하는데, 데이터가 heavily censored된 상황이나 불량이 매우 적은 상황에서의 logistic regression 등에 활용은 어떻게 할 수 있는지요?

- For bootstrapping censored data, see Efron (1981).
- How to incorporate in logistic regression?
  I do not know.

References

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, UK.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, 76:312–319.

Gumbel, E. J. (1963). Statistical forecast of droughts. *Bulletin – International Association of Scientific Hydrology*, 8:5–23.

Hettmansperger, T. P. and McKean, J. W. (2010). *Robust Nonparametric Statistical Methods*. Chapman & Hall/CRC, Boca Raton, FL, 2nd edition.

Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, 5:201–213.

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.

References

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the 1975 International Congress of Mathematics, vol. 2*, pages 523–531. Vancouver.