Miscellaneous

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

Overview

A measure of heaviness of a tail

Definition of kurtosis

$$\text{kurtosis} = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right].$$ Note: the kurtosis of normal is **3**.

NIST: kurtosis

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with **high kurtosis tend to have heavy tails, or outliers**. See NIST (2018) at ▸ NIST.

Interpretation of kurtosis (Wikipedia)

$\cdots$ The exact interpretation of the measure of kurtosis used to be disputed, but is *now settled(?)* (Westfall, 2014). $\cdots$ Therefore, kurtosis measures outliers only; it measures nothing about the peak(?).
See ▸ wiki(Kurtosis).

尖度 (일본) 峯度 (중국) 鈍尖度 (my suggestion)

A measure of heaviness of a tail

Heavy distribution

- Rigorously, a heavy-tailed distribution means any distribution whose tail has heavier tail(s) than the **exponential**.
  See ▸ wiki(heavy tail).
- **Occasionally**, a heavy-tailed (fat-tailed) distribution means any distribution that has **heavier tail(s) than the normal**.
  (kurtosis greater than 3).

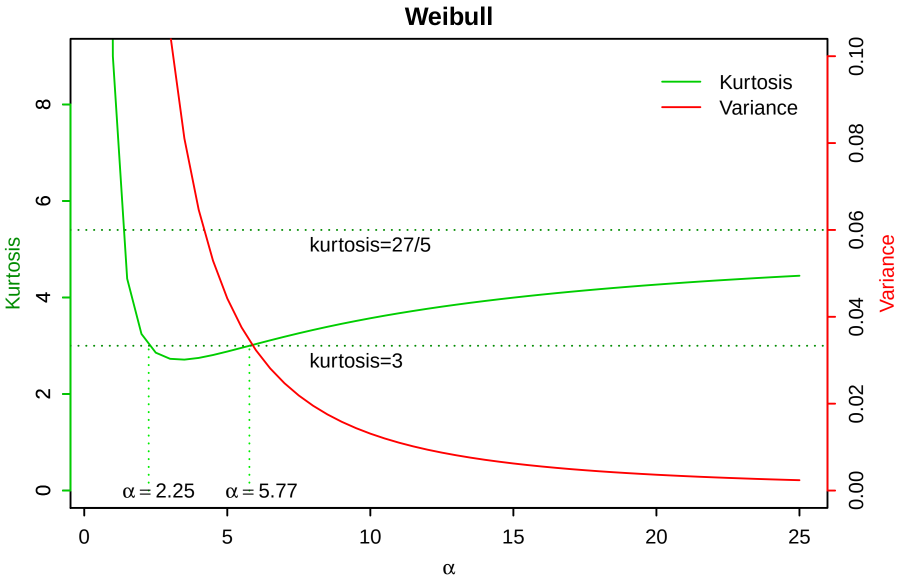
Danger of using kurtosis (Example: Weibull distribution)

$$\text{kurtosis} = \frac{\Gamma(1+\frac{4}{\alpha}) - 4\Gamma(1+\frac{3}{\alpha})\Gamma(1+\frac{1}{\alpha}) + 6\Gamma(1+\frac{2}{\alpha})\Gamma(1+\frac{1}{\alpha})^2 - 3\Gamma(1+\frac{1}{\alpha})^4}{\left(\Gamma(1+\frac{2}{\alpha}) - \Gamma(1+\frac{1}{\alpha})^2\right)^2}$$

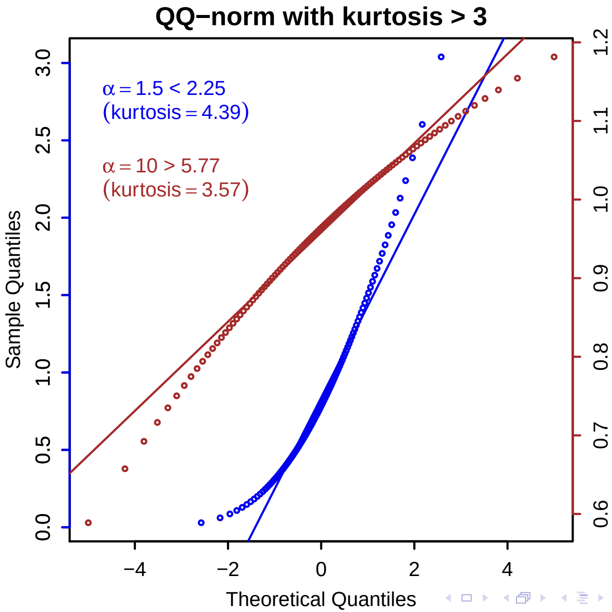
$$\text{variance} = \Gamma(1+\tfrac{2}{\alpha}) - \Gamma(1+\tfrac{1}{\alpha})^2$$

$\Diamond$NB: as $\alpha \to \infty$, the variance goes to 0 and the kurtosis goes to $27/5$.

A measure of heaviness of a tail (danger of using kurtosis)



A measure of heaviness of a tail (danger of using kurtosis)



A measure of heaviness of a tail (alternative to kurtosis)

- In some cases, the kurtosis does not explain the heaviness of a distribution properly.

  $\Longrightarrow$ Use the QQ-norm if one wants to compare with a normal distribution. (**baseline**: normal distribution)

  $\Longrightarrow$ Use the definition of heavy-tailed dist. provided in Foss et al. (2013). (**baseline**: exponential distribution)
  According to this definition, Weibull has a heavy tail only if α (shape) less than one.

- Also, for many distributions including the Cauchy, t-distributions with 1,2,3,4 degrees of freedom, and so on, the value of the kurtosis is infinity (∞).

  $\Longrightarrow$ The definition in Foss et al. (2013) can be used, but it does not provide the strength of heaviness.

  $\Longrightarrow$ L-kurtosis by Hosking (1990) can be a good alternative.

$\Diamond$ Suggestion: **Kurtosis** + **QQ-norm** + L**-kurtosis**.

$\Diamond$ Not a fan of Foss et al. (2013): too complex and no strength.

Different meanings of robustness in engineering

- In practice, robust estimators are widely used when dealing with **heavy-tailed** distributions and **contaminated** data, etc.
- We will look at various interpretations of robustness.

Definition of Robustness in statistics literature

- Robustness implies **stability** of parameter estimation under departure from the true model (contamination, model departure, etc.) In statistics, it is also called **outlier-resistance**.
- Outliers can be from contamination or from a nature of **heavy** distribution (surprising observation).

◇ The meaning of robustness in engineering is slightly different.

Different meanings of robustness in engineering

Robust to noise. Noise in what sense?

- Robust to **contamination** (outlier / influential).
  Robust to something wrong.
- Robust to **model departure** (usually departure from the normality).
  Robust to something different from normal.
  For example, the t-test is robust; see Remark 8.3.1 of (Hogg et al., 2013). (Roughly speaking, it is due to CLT).
- Robust to **surprising observation**
  Robust to something surprising.
  An outlier (from a **heavy** distribution). Thus, not a contamination.
  Not necessarily influential.
  No influential observation to the MLE of Laplace (median).
  For Cauchy, we can easily meet surprising outlying observations.
- Robust to **uncontrollable noise** (Robust Design)
  Robust to something uncontrollable.

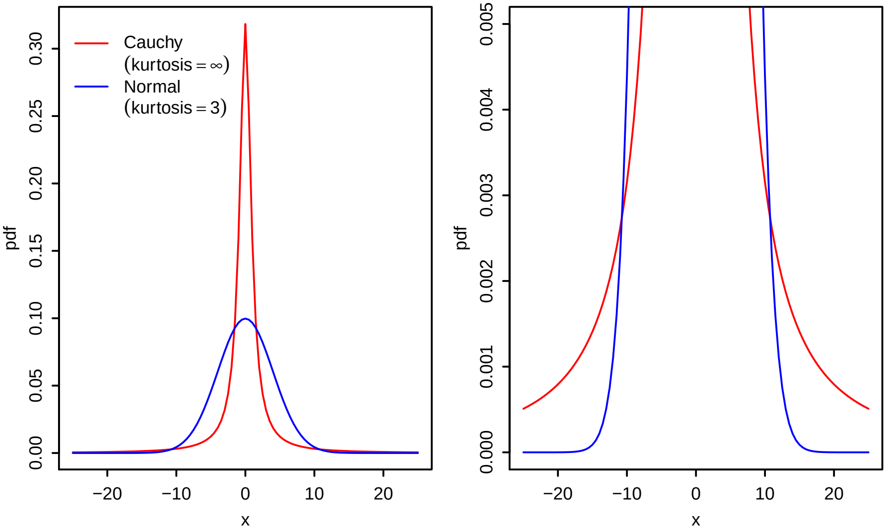
Measures of robustness property (review)

- Robust to **contamination** $\Longrightarrow$ **Breakdown point**.
  The percentage of contaminated observations before the estimator becomes non-sensical and incorrect.
  Refer to $\boxed{\text{Talk-1}}$ at ▸ Seminar/2018.
- Robust to **model departure** $\Longrightarrow$ **Relative efficiency**.

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

  where $\hat{\theta}_1$ is often a reference or baseline estimator (usually, the MLE under a normal distribution).
  Refer to $\boxed{\text{Talk-5}}$ at ▸ Seminar/2018.
- Robust to **uncontrollable noise** (Robust Design) $\Longrightarrow$ **SN ratio**.

Cauchy: $f(x) = \frac{\beta}{\pi} \frac{1}{\beta^2 + (x-\alpha)^2}$ Normal: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$



Cauchy and Normal distribution

R Code to compare Cauchy and Normal

```
set.seed(14)  # Change this seed.
n = 10;       # n = 10000
x = rcauchy(n)
y = rnorm(n)
OUT = rbind( c(mean(x),sd(x)), c(mean(y),sd(y)) )
colnames(OUT) = c("mean",  "SD")
rownames(OUT) = c("Cauchy(x)","Normal(y)")
OUT
```

- Cauchy is very dangerous due to **gravity with leverage**.
- Robust estimators are needed for most of heavy-tailed distributions.
- MLE under heavy distributions are robust in general.
  (MLE of t-distribution, Laplace, Cauchy, etc).

Smooth empirical distribution and KDE

Suppose that there are n observations: $x_1, x_2, \ldots, x_n$, and their order statistics are denoted by $y_1 < y_2 < \ldots < y_n$. The value of the empirical distribution at a given point, y_k, is given by

$$\hat{F}_n(y_k) = \frac{k}{n}.$$

In general, the empirical distribution function of t (**without sorting**) can be written by

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i \le t),$$

where $\mathbb{I}(\cdot)$ is an indicator function.

$\Diamond$NB: $\hat{F}_n(t)$ also works for tied observations.

We can re-write $\hat{F}_n(t)$ as

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(X \le t | \mu_i = x_i),$$

where X is a random variable which degenerates at $\mu_i = x_i$.
Instead of degenerating $\mathbb{P}$, one can use any smooth distribution function.
Say, we use normal CDF with $\mu_i = x_i$ and $\sigma_i = h_i$ (aka, bandwidth). That is, we have

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} P(X \le t \mid \mu_i = x_i, \sigma_i = h_i),$$

where $X \sim N(x_i, h^2)$. Since $Z = (X - \mu_i)/\sigma_i$, we have

$$\begin{aligned}\hat{F}_n(t) &= \frac{1}{n}\sum_{i=1}^{n} P\Big(Z \le \frac{t-\mu_i}{\sigma_i} \mid \mu_i = x_i, \sigma_i = h_i\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n} P\Big(Z \le \frac{t-x_i}{h_i}\Big) = \frac{1}{n}\sum_{i=1}^{n} \Phi\Big(\frac{t-x_i}{h_i}\Big), \qquad (1)\end{aligned}$$

where Φ is a CDF of $N(0,1)$.

Differentiating Eq. (1) with t, we have

$$\hat{f}_n(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i}\phi\Big(\frac{t-x_i}{h_i}\Big),$$

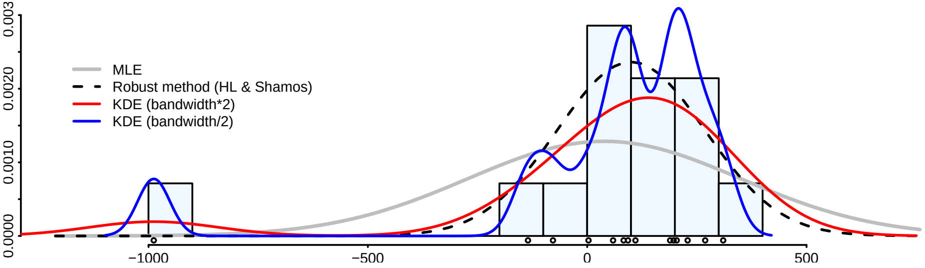
Instead of $\phi(\cdot)$, one can use any kernel $k(\cdot)$ which is positive and its integration is one. Also, we can use any positive weights instead of 1. Then we have more general KDE

$$\hat{f}_n(t) = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i\frac{1}{h_i}k\Big(\frac{t-x_i}{h_i}\Big).$$

In many cases, we use a constant weight $w_i = 1$ and a fixed bandwidth $h_i = h$. Then we have

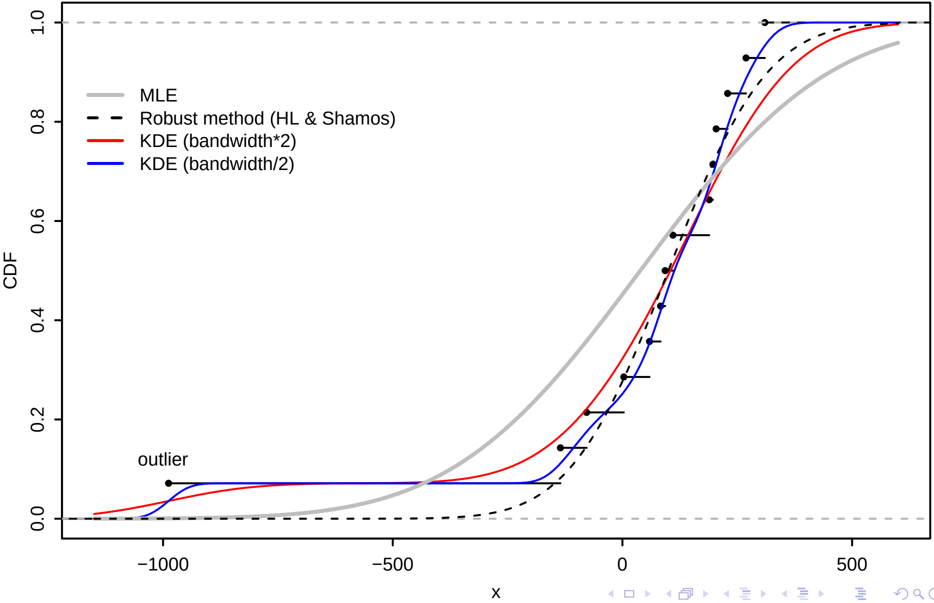
$$\boxed{\hat{f}_n(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k\Big(\frac{t-x_i}{h}\Big) = \frac{1}{nh}\sum_{i=1}^{n}k\Big(\frac{t-x_i}{h}\Big).}$$

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

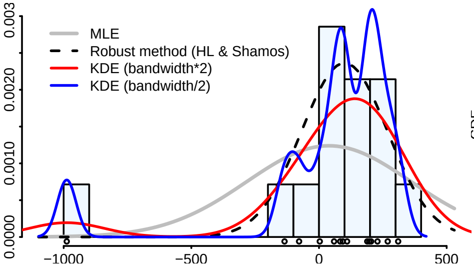


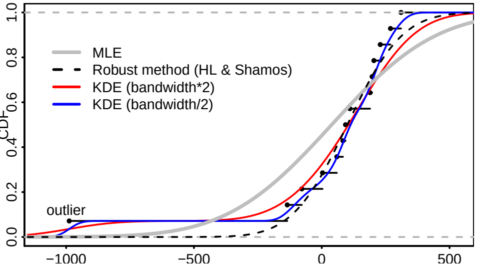
- $-988, -135, -78, 3, 59, 83, 93, 110, 189, 197, 204, 229, 269, 310.$
- Using the empirical CDF, $\hat{F}_n(-500) = 1/14 = 0.07142857$. ($n = 14$).
- Using the CDF with the MLE, $\hat{F}_{\hat{\theta}}(-500) = 0.04706641$.
- Using the CDF with robust estimation, $\hat{F}_{\hat{\theta}}(-500) = 0.0001891299$.
- Using the CDF with the KDE (large h), $\hat{F}_h(-500) = 0.07203472$.
- Using the CDF with the KDE (small h), $\hat{F}_h(-500) = 0.07142857$.

Revisit: robustness issues (CDF)



Revisit: robustness issues (pdf and KDE)





- The pdf (grey) with the MLE does not fit the histogram. ($\hat{\mu} = 38.92857$ and $\hat{\sigma} = 321.9428$). NB: gravity and leverage.
- The pdf (dashed) with the HL and Shamos fits the histogram well. ($\hat{\mu} = 100.0$ and $\hat{\sigma} = 168.7856$). NB: it is robust.
- The KDE (blue/red) is not robust. But, the probability estimation is not so sensitive to an outlier.
- The empirical CDF (step function) is not robust either. But, the probability estimation is not so sensitive to an outlier.
- The KDE depends on the bandwidth h.

Skewed distribution

Skewness

- Skewness is a measure of the **asymmetry** of a pdf.
  (zero $\Longrightarrow$ symmetric. non-zero $\Longrightarrow$ asymmetric or skewed).
- A unimodal left-skewed distribution has a long left tail while a unimodal right-skewed distribution has a long right tail.
- A skewed distribution with a light longer tail is **not lethal**.
- A **long and heavy** tail may create potential problems.

Measure of skewness

- $E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ (Pearson). Most popular measure.
- $\frac{\boldsymbol{\mu}-\text{median}}{\sigma}$ (old **nonparametric**).
  $\Diamond$NB: $\boldsymbol{\mu}$ is just a center of gravity (so nonparametric).
- Others: quartile-based measure, L-moment by Hosking (1990), etc.

Skewed distribution

How to deal with skewed distributions

- One can use the empirical CDF or the KDE although they may not be robust.
  Note: the estimation of a probability is not so sensitive to outliers.
- One can select parametric (asymmetric or skewed) distributions with robust estimation.
  For example, Weibull, lognormal, Birnbaum-Saunders, gamma distribution, χ^2-distribution, F-distribution, etc.
- One can normalize the data set (say, using Box–Cox Transformation). Then, use the existing methods developed under the normality.
- One can try data-driven weights for a general KDE to give smaller weights for outliers

$$\hat{f}_n(t) = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \frac{1}{h} k\left(\frac{t - x_i}{h}\right).$$

References

Foss, S., Korshunov, D., and Zachary, S. (2013). *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer, New York.

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

Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society B*, 52:105–124.

NIST (2018). NIST/SEMATECH e-handbook of statistical methods: 1.3.5.11 measures of skewness and kurtosis. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm. Accessed: 2018-08-07.

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

Westfall, P. H. (2014). Kurtosis and peakedness, 1095–2014. R.I.P. *The American Statistician*, 68:191–195.