# Density Estimation

Udayan Nath(191168)
Vinay Kumar Sharma(191171)
Vishnudatt Nagar(181172)

Department of Mathematics and Statistics , IIT KANPUR

April 18, 2021

# Introduction

- In statistics, kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable.
- Let $(x_1, x_2, \ldots, x_n)$ be independent and identically distributed samples drawn from some univariate distribution with an unknown density $f$ .
- At any given point $x$ its kernel density estimator is

$$f_h\hat{(}x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

where,$K$ is the kernel function, h is the bandwidth $(h = ((\frac{4\hat{\sigma}^5}{3n})^{\frac{1}{5}})$,$\hat{\sigma}$ is sample standard deviation and $x_i$ , i=1(1)n is the transformed data.

# Motivation

- Kernel density estimation perform well if the density is Gaussian or approxmiately Gaussian in shape.
- But if the distribution is heavy tail then usual estimation method performs poorly, which can be solved using transformations
- In this project work we performed simulation studies to show the problems of using the untransformed density estimation for highly skewed data and how using transformation can help us to overcome the problem. Then we used the method to analyze real life data sets of car insurance.

# Transformations

- There are different types of transformations available like Shifted Power Transformation( by Wand et al), Champernowne Transformation( by Buch-Larsen et al) and Mobius-like transformation (by Clements et al).
- But we used the Möbius like transformation, scale parameter R and $\alpha$ for highly skewed data.

# Möbius like transformation

**Mobius-like transformation (by Clements et al)**

$$T_{\alpha,R}(x) = \frac{x^\alpha - R^\alpha}{x^\alpha + R^\alpha}, \forall x \in [0, \infty] \tag{1}$$

where $\alpha > 0$, $R > 0$,

- The probability density function $f(x)$ of X and $f(z)$ of Z are connected by the formula,

$$f_X(x) = f_Z(z)\frac{2\alpha R^\alpha x^{\alpha-1}}{(x^\alpha + R^\alpha)^2} = \frac{\alpha}{2R} f_Z(z)(1+z)^{1-\frac{1}{\alpha}}(1-z)^{1+\frac{1}{\alpha}} \tag{2}$$

# Asymptotic Theory for Transformation

- For the $x_1, x_2, \ldots, x_n$ transformed density estimator

$$f(\hat{x}) = \frac{1}{n} \sum_{i=1}^{n} (K_h(T(x) - T_i(x)) T^{'}(x)$$

  where, T(.) is the transformation function, $T^{'}$ is the first differential and $K_h$ is the kernel function with bandwidth h.

- Now, the bias and the variance of $f(x)$ are given by,

$$Bias = E[f(\hat{x})] - f(X) = \frac{1}{2}\mu_2(K)h^2((\frac{f(x)}{T'(x)})\frac{1}{T'(x)}) + 0(h^2)$$

$$Variance = V(f(\hat{x})) = \frac{1}{nh} R(K) T^{'}(x) f(x) + 0(\frac{1}{nh})$$

# Asymptotic Theory for Transformation Continued...

where $\mu_2(K) = \int(u^2K(u)du$ and $R(K) = \int(K^2(u)du$.

- Now, as for the classical density estimation the classical density estimator follows a normal distribution asymptotically as $n \longrightarrow \infty$:

$$\sqrt(nh)(g\hat(z) - E(g\hat(z)) \sim N(0, \frac{1}{nh}R(K)g(y))$$

  then as $\hat{f}(x) = T^{'}\hat(x)g(x)$ with $z = T(x)$. Then

$$\sqrt(nh)(f\hat(x) - E(f\hat(x)) \sim N(0, \frac{1}{nh}R(K)T^{'}(x)f(x))$$

  where $z_i = T(x_i)$ i.e. transformed variable have distribution $g$ and $g\hat(z)$ is the classical kernel density estimate of $g(z)$.

- The proofs regarding the asymptotic theory of classical kernel density estimator is done in the book by E.L.Lehmann and the proof for the transformed kernel density estimate is in the paper by Buch-Larsen et al.

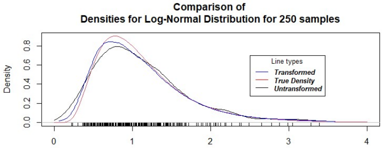
# Simulation Study

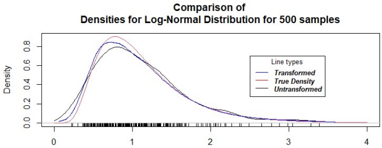
- In this section a comparison between untransformed and transformed kernel density estimates is made. We used two different heavy-tailed distributions with different tails and shape to simulate data.
- We used Weibull and Lognormal distribution to simulate data, both of which are heavy -tailed in nature.
- The distribution and chosen parameters are listed below:

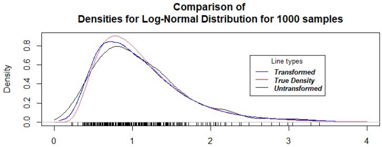
| Distribution | Density for x>0 | Parameters |
|---|---|---|
| Lognormal($\mu,\sigma^2$) | $f(x)= \frac{1}{\sqrt{2\pi\sigma^2 x}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$ | $(\mu,\sigma^2)=(0,0.5)$ |
| Weibull($\gamma$) | $f(x)= \gamma x^{(\gamma-1)} e^{-x^{\gamma}}$ | $\gamma= 1.5$ |

# Simulation Continued...

## Lognormal distribution

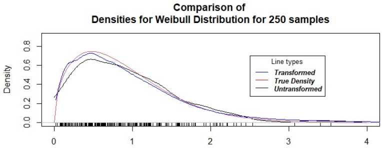


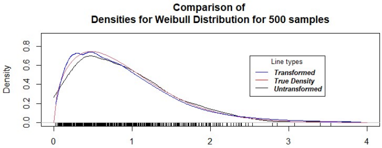



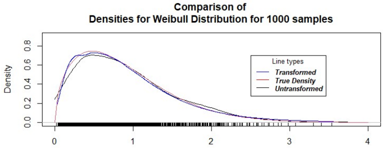



# Simulation Continued...

## Weibull distribution



# Performance Measure

- The Integrated squared error (ISE) is defined as,

$$ISE = \int_0^{\infty} (f_h\hat{(}x) - f(x))^2 \, dx$$

- The Weighted integrated squared error (WISE) is defined as,

$$WISE = \int_0^{\infty} (f_h\hat{(}x) - f(x))^2 x^2 \, dx$$

- The mathematical expression of Kullback-Leibler Divergence ,for Continuous probability distribution $P$ and $Q$

$$D_{kl} = \int_{x \in \chi} P(x) \log(\frac{P(x)}{Q(x)}) dx$$

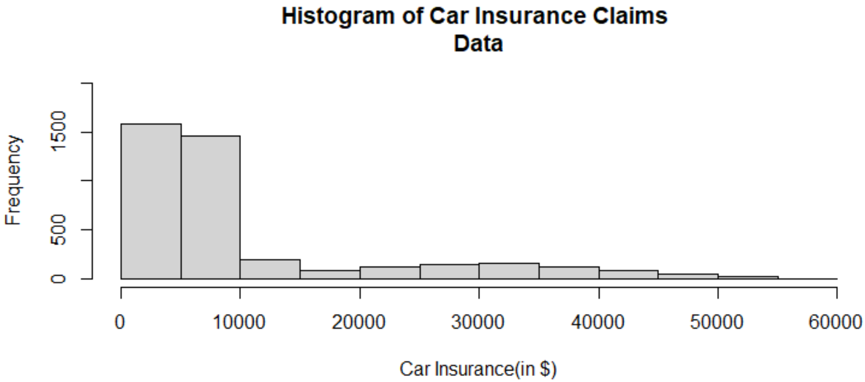
# Comparison

- Here we compare Untransformed and Transformed kernel density estimator using the values of ISE , WISE and KL divergence for two different distributions i.e Log normal and Weibull.

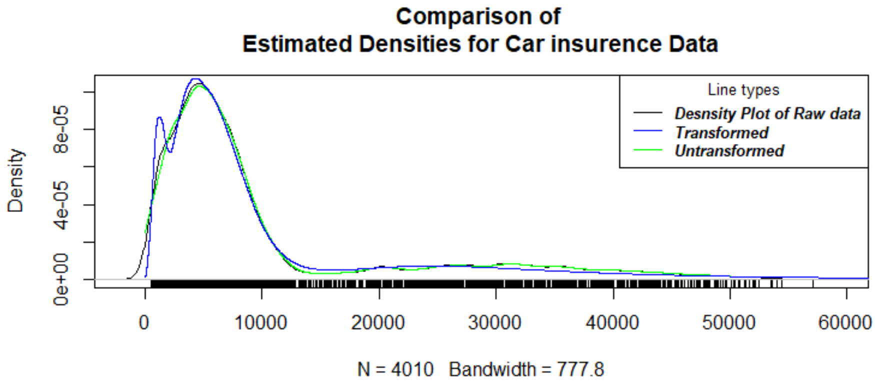
| Sample size | Error | PROBABLITY DENSITY | | | |
|---|---|---|---|---|---|
| | | Weibull | | Log Normal | |
| | | Untransformed | Transformed | Untransformed | Transformed |
| n = 250 | ISE | 0.2828 | 0.0022 | 0.3373 | 0.0036 |
| | WISE | 0.2744 | 0.0043 | 0.3675 | 0.0020 |
| | KL | 0.0799 | 0.5072 | 0.0170 | 0.3742 |
| n = 500 | ISE | 0.2745 | 0.0015 | 0.3373 | 0.0036 |
| | WISE | 0.2756 | 0.0021 | 0.3675 | 0.0020 |
| | KL | 0.0844 | 0.5209 | 0.0171 | 0.3750 |
| n = 1000 | ISE | 0.2651 | 0.0009 | 0.3374 | 0.0036 |
| | WISE | 0.2721 | 0.0014 | 0.3676 | 0.0021 |
| | KL | 0.0832 | 0.5312 | 0.0171 | 0.3752 |

# Data Analysis

- We do our study on Car-Insurance Claims Data and the histogram and estimated desity plot of transformed and untransformed data is given below.



# Data Analysis Continued...



# Conclusion

- From the simulated study we observed that, in each case for highly skewed data set it is convenient to use transformation.
- It helps to reduce the error as the transformed estimator estimates the tail.
- The value of KL divergence for untransformed and transformed data is respectively 0.000002537 and 0.000102. which is significantly much smaller. So, we conclude that for this highly skewed real life data transformation gives better estimation of density.

*Thank You*