Lecture 11 – Gaussian and transformation

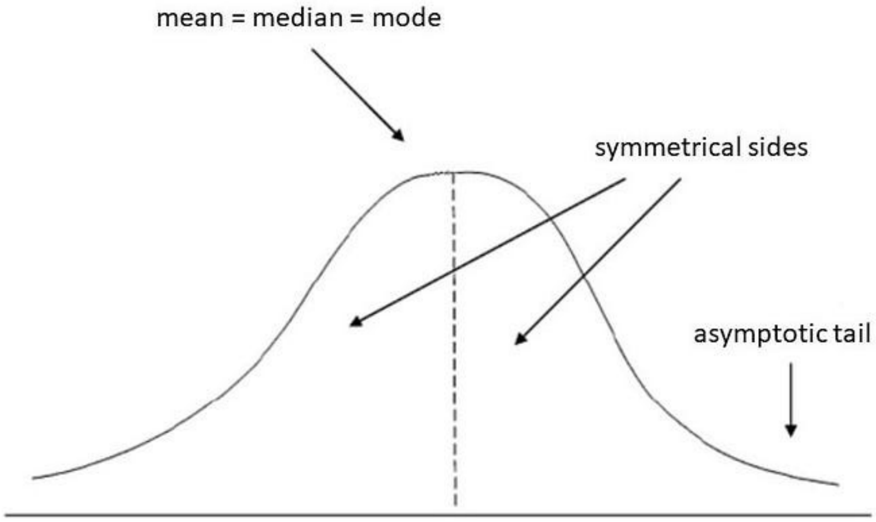
Wengang Mao (Marine Technology)
Department of Mechanics and Maritime Sciences,
Chalmers University of Technology,
Goteborg, Sweden

CHALMERS
UNIVERSITY OF TECHNOLOGY

Outline

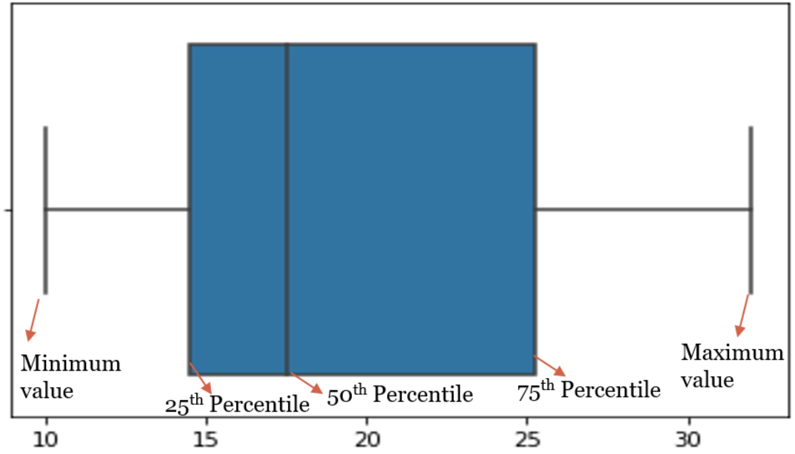
- Background of Gaussian
  - ✓What is Gaussian; how to check Gaussian
  - ✓Why Gaussian
- Techniques for the transformation
  - ✓Log Transformation
  - ✓Reciprocal Transformation
  - ✓Exponential Transformation
  - ✓Box-Cox Transformation
  - ✓Normal Inverse Gaussian

What is Gaussian

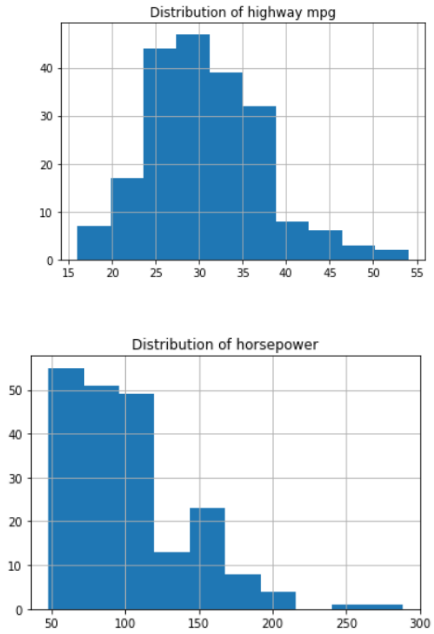


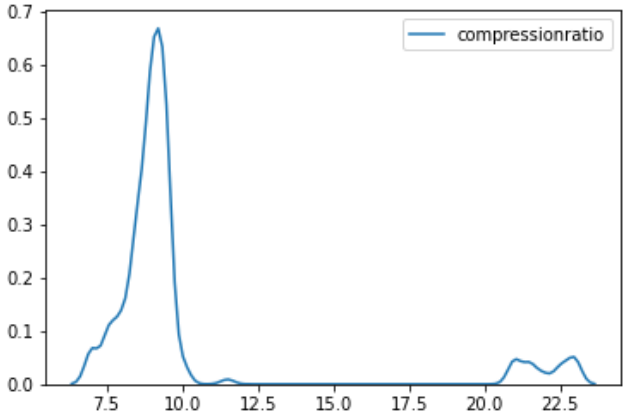
How to check Gaussian

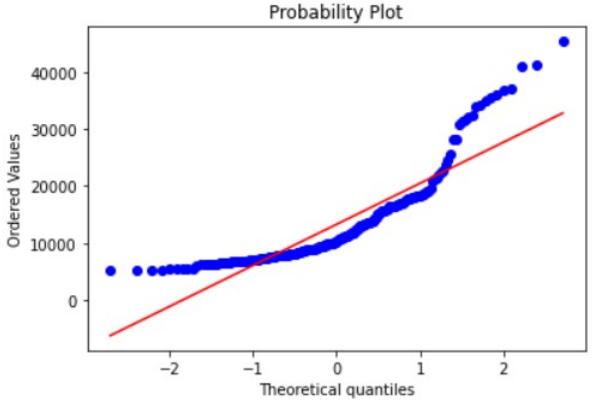
- Five Number Summary
- Histogram
- Q-Q plot
- KDE plot
- Skewness and Kurtosis



How to check Gaussian







Why Gaussian

- **Importance of Normality in Machine Learning!**
  1. Gaussian distribution is found everywhere because a dataset with finite variance turns into Gaussian as long as the dataset with independent feature-probabilities is allowed to grow in size.
  2. Datasets with Gaussian distributions make applicable to a variety of methods that fall under parametric statistics.
     The methods such as propagation of uncertainty and least squares parameter fitting that make a data-scientist life easy apply only to datasets with normal or normal-like distributions.
  3. Since Gaussian(Normal) Distribution is easy to explain, the intuition behind the conclusion and summary of the test can be easily conveyed to people with little statistical knowledge.
  4. The entire distribution is described by two numbers, the mean and variance.
  5. Unlike much other distribution that changes their nature on transformation, a Gaussian tends to remain a Gaussian.

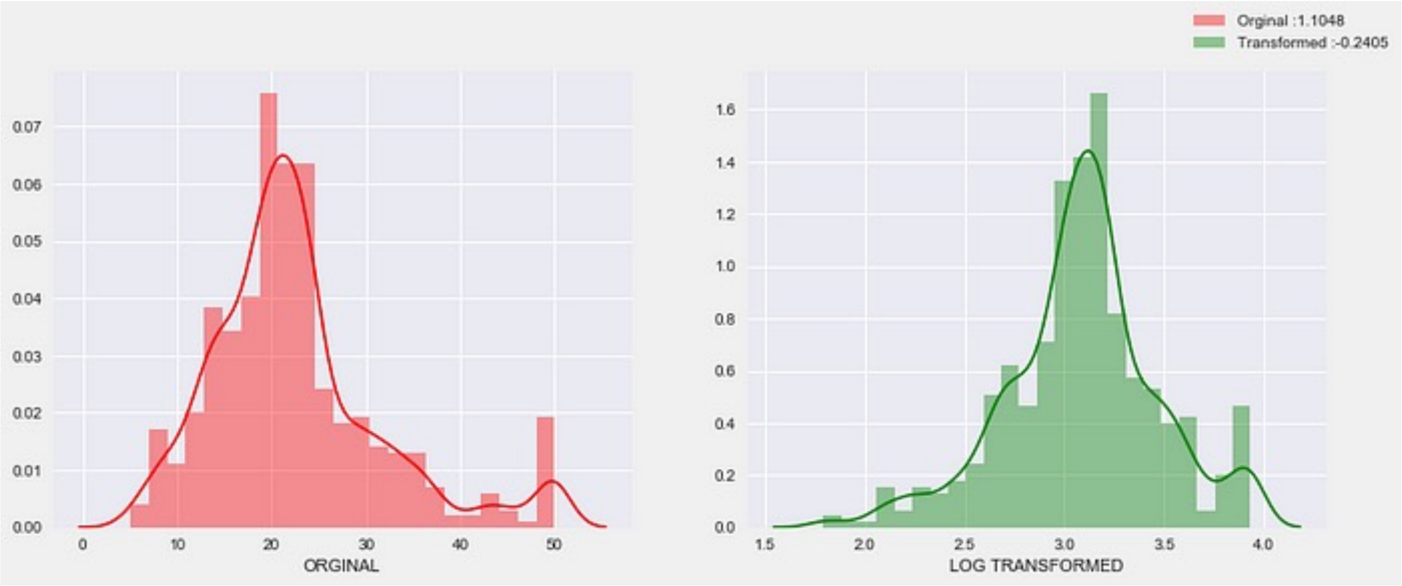
     * The product of two Gaussian is a Gaussian
     * The Sum of two independent Gaussian random variables is a Gaussian
     * Convolution of Gaussian with another Gaussian is a Gaussian
     * Fourier transform of Gaussian is a Gaussian

Need for Data Transformation!!

- To more closely approximate a theoretical distribution that has nice statistical properties
- To spread data out more evenly -to make data distributions more symmetrical
- To make relationships between variables more linear
- To make data have more constant variance (homoscedastic)

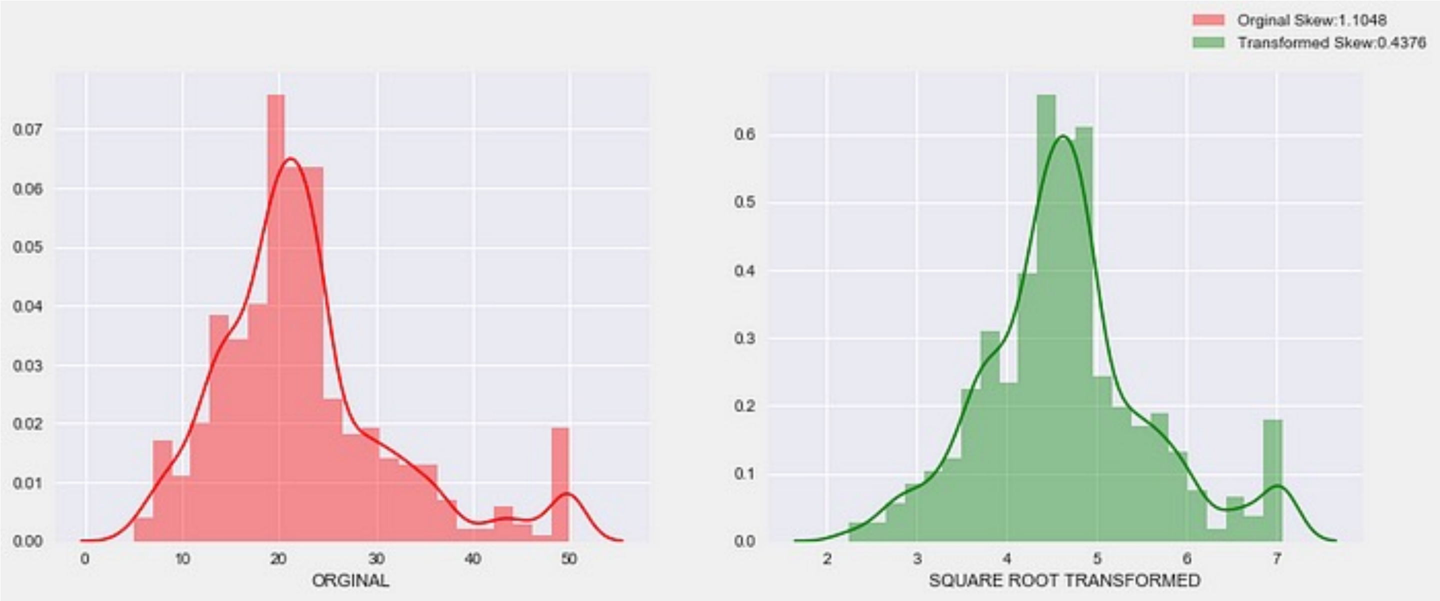
Transformation (1) -- Log Transformation

- Y = log(X)



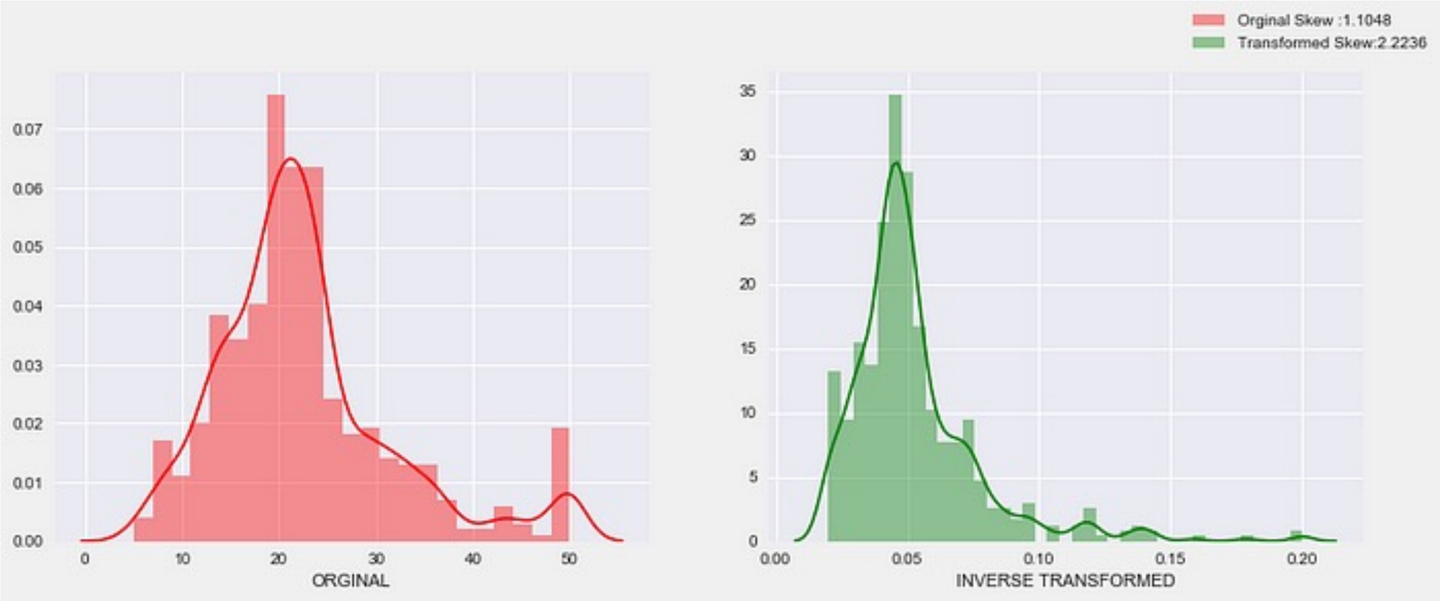
Transformation (2) -- Exponential Transformation

$Y = \exp(X)$; or $Y=X^n$

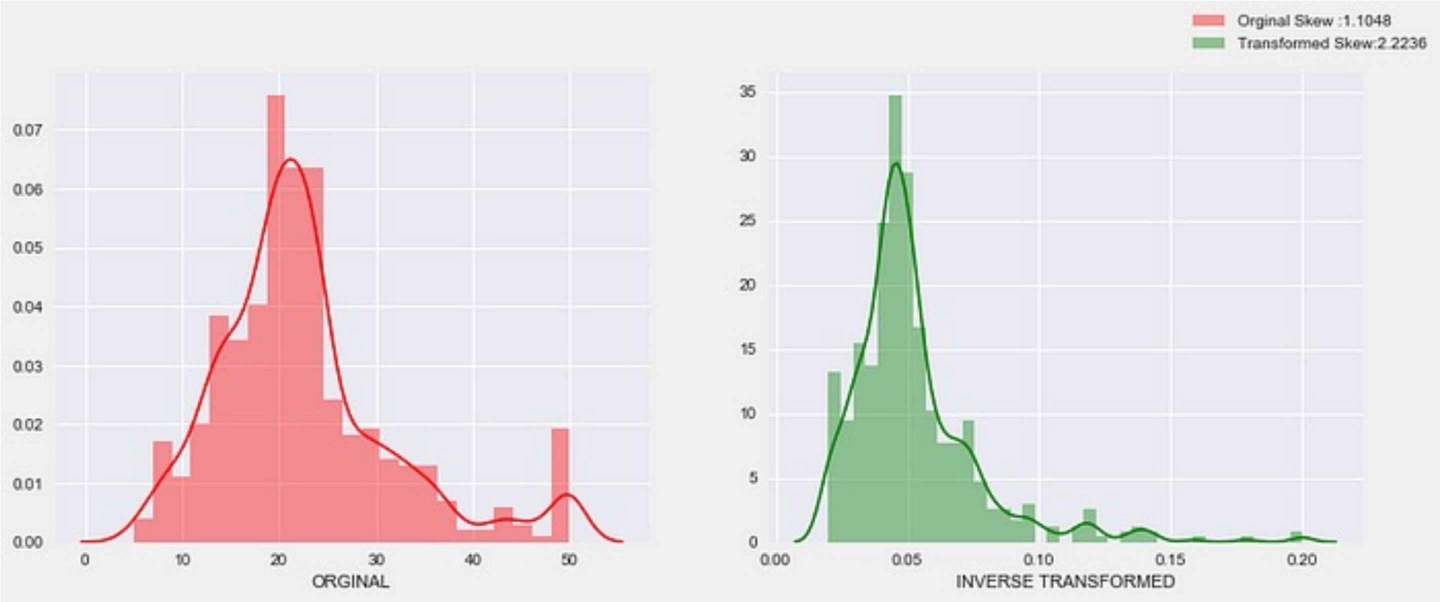


Transformation (3) -- Reciprocal Transformation

$Y = 1/X$



Transformation (4) -- Reciprocal Transformation



Normal Inverse Gaussian

Normal inverse Gaussian distribution

- A relatively flexible probability distribution that can be used to approximate unknown probability distributions by fitting it to data.
- Can fit mean, variance, skewness, and kurtosis to data.
- We can also extend it to a *log-NIG distribution* to make it even more flexible.
- After fitting, data can be mapped to (marginal) Gaussianity.

Normal-inverse Gaussian (NIG)

Parameters	μ location (real) α tail heaviness (real) β asymmetry parameter (real) δ scale parameter (real) $\gamma = \sqrt{\alpha^2 - \beta^2}$
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{\alpha\delta K_1\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)}{\pi\sqrt{\delta^2+(x-\mu)^2}}\, e^{\delta\gamma+\beta(x-\mu)}$ K_j denotes a modified Bessel function of the second kind[1]
Mean	$\mu + \delta\beta/\gamma$
Variance	$\delta\alpha^2/\gamma^3$
Skewness	$3\beta/(\alpha\sqrt{\delta\gamma})$
Ex. kurtosis	$3(1+4\beta^2/\alpha^2)/(\delta\gamma)$
MGF	$e^{\mu z+\delta(\gamma-\sqrt{\alpha^2-(\beta+z)^2})}$
CF	$e^{i\mu z+\delta(\gamma-\sqrt{\alpha^2-(\beta+iz)^2})}$

AVANCEZ

1829

CHALMERS

UNIVERSITY OF TECHNOLOGY