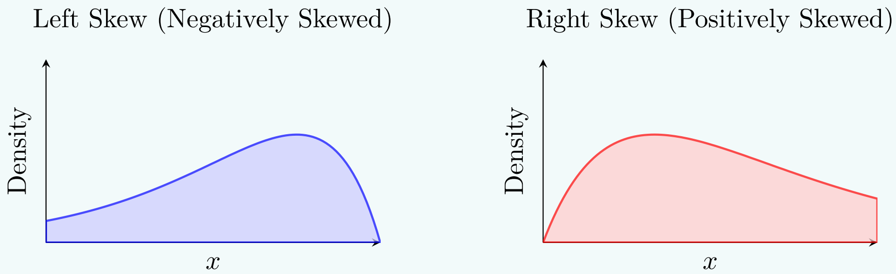
STAT 151A Lecture 3

Henry Liev

3 September 2025

Method 0.1 (Transforming to fix skewness)



Will want to compress distances between values log will compress for right skew and e^x will compress for left skew
To pick a specific transform $f(x)$ use $\frac{f(\text{Upper Quartile})-f(\text{Median})}{f(\text{Median})-f(\text{Lower Quartile})} = 1$

Example 0.2 (Kleiber Data Box Cox Transformation)

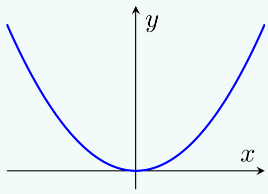
Transform	$\frac{f(\text{UQ})-f(\text{Med})}{f(\text{Med})-f(\text{LQ})}$
x	2.5
$\sqrt{x}$	1.23
$\sqrt[3]{x}$	0.96
log	0.58

Remark 0.3 (Transforming Nonlinear Relationships)

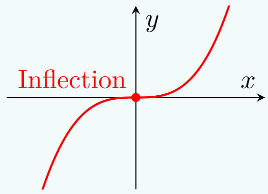
2 things we need:

1. "Simple" - direction of curvature (2nd derivative) does not change

Simple: Quadratic

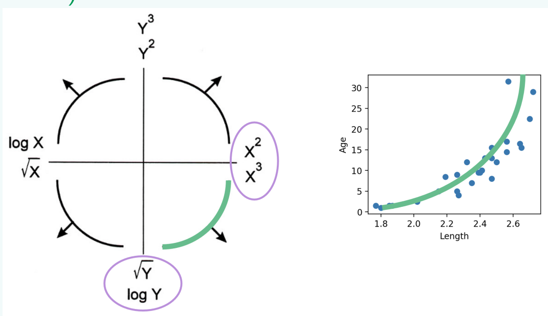


Not Simple: Cubic



2. Monotone - strictly increasing or decreasing

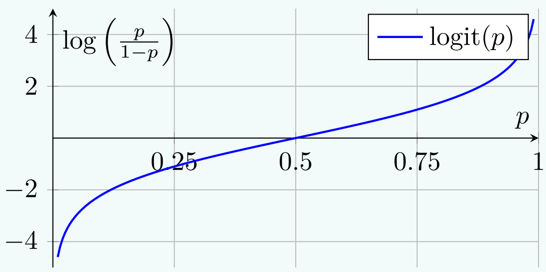
Remark 0.4 (Transformations)



Transform y or transform x
What about data where $y \in (0, 1)$?
Issues:

- Scale, power transformations decrease instead of increase, negative values, all values are huge (rescale)
- Density (Binary) apply logit$(p) \to \log(\frac{p}{1-p})$

Logit Function



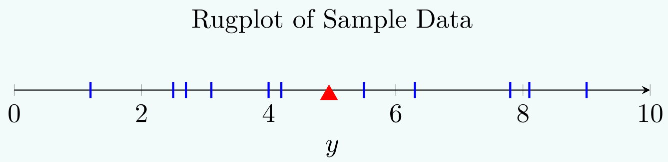
Caution: Not always necessary and hurts interpretability

Remark 0.5 (Revisit mean and SD using regression)

We can think of the mean, sd, and simple regression in 4 forms: Graphically, Optimization, Probabilistically, Vector Spaces $y_1, \dots, y_n$

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, SD = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$

Graphically, the mean is the "balance point" of these plots, SD: How spread out are the points



Optimization

$$\min_c \sum_{i=1}^{n}(y_i - c)^2 \to \frac{\partial}{\partial c}\sum_{i=1}^{n}(y_i - c)^2 = 0 \to \hat{c} = \bar{y}$$

$\bar{y}$ is the minimizer of the squared error

$$\sum(y_i - c)^2 = \sum(y_i - \bar{y} + \bar{y} - c)^2 = \sum(y_i - \bar{y})^2 + \sum(c - \bar{y})^2$$

SD? Minimized objective value is $\sum(y_i - \bar{y})^2 = (n-1)s^2$ SD tells us how good of a miniizer $\bar{y}$ is

Probability

RV $Y_i = \mu + \varepsilon_i$ where $\mathbb{E}(\varepsilon_i) = 0, \mathrm{Var}(\varepsilon_i) = \sigma^2$ and μ is a parameter

If $y_1, \dots, y_n$ are random variables from the random variable model, then can we use them to estmiate μ?

Yes and $\bar{y}$ is a "good" estimator of μ

What about σ^2? s^2 is a good estimator for σ^2