Data transformation and standardization
UNIT 3

Dr. Praveen Barapatre

Freeman-Tukey transformation

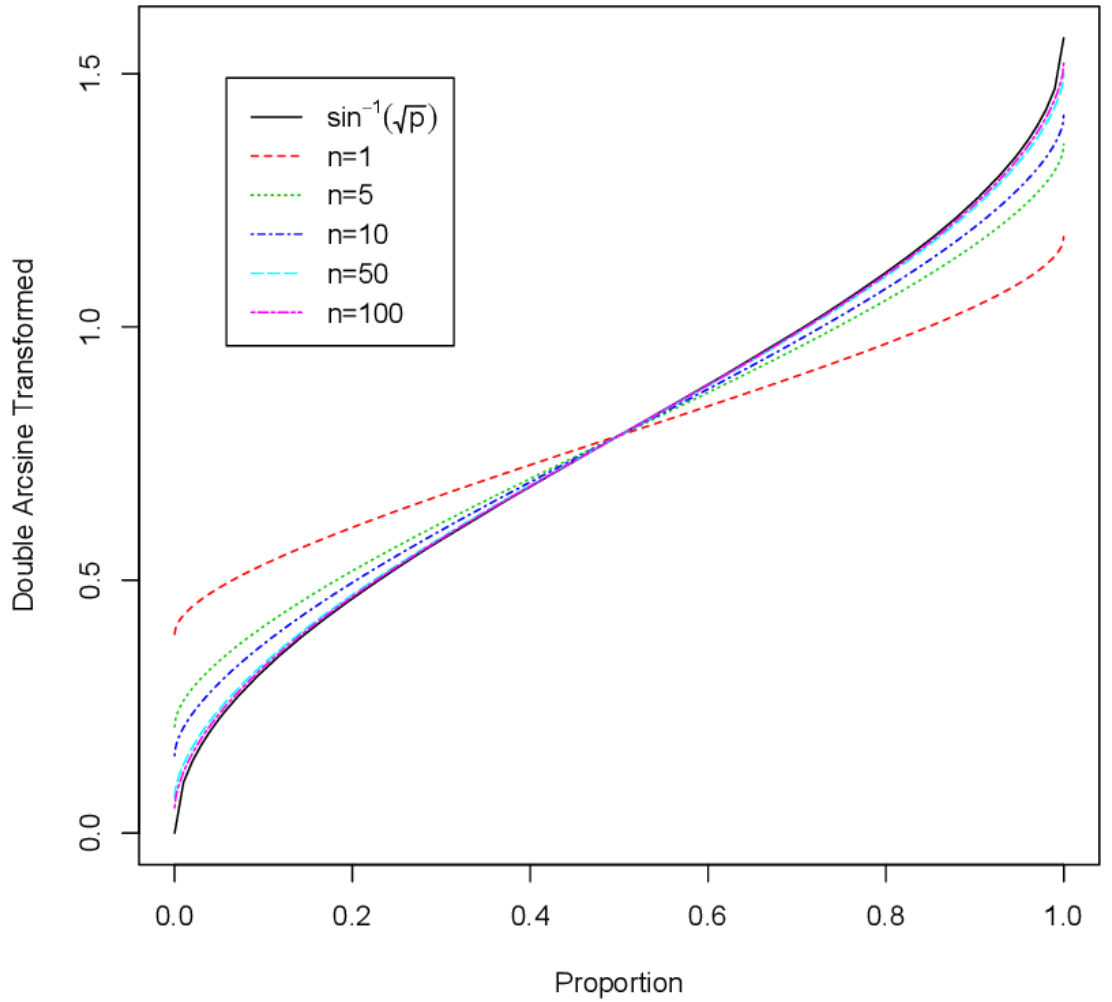
- The Freeman-Tukey transformation is a data transformation technique used to normalize proportions for statistical analysis.
- It is particularly useful when working with proportions that come from small samples or when the proportions are close to 0 or 1.
- The transformation makes the data more closely follow a normal distribution, which allows for the use of standard statistical tests.

Two main applications of the Freeman-Tukey transformation

Transforming proportions:

- If X is a binomial variable with parameters n (number of trials) and p (probability of success), then the transformed variable Y, given in radians by the following formula, has an approximate normal distribution with mean sin^-1(√p) and variance 1/(4n + 2).

Graph of double arcsine transformation for different sample sizes; sin−1(√ p) is the limiting function as n→∞.



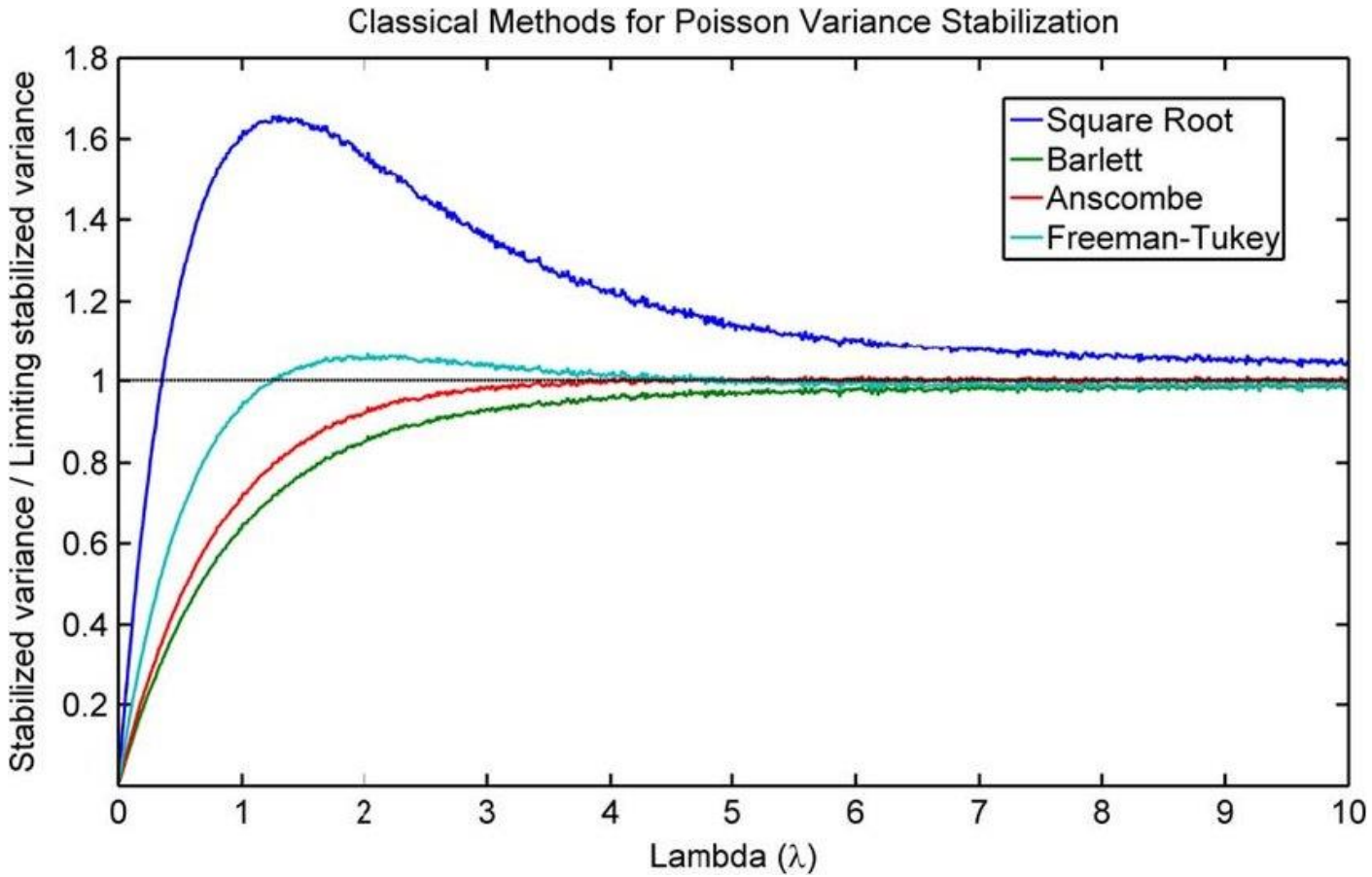
The approximation improves as np (number of successes) increases and should not be used if np < 1.

Two main applications of the Freeman-Tukey transformation

Transforming Poisson variables:

If X is a Poisson variable with expectation μ (average number of events), then, for μ>1, the transformed variable Z, given by the following formula, has an approximate standard normal distribution. Observed values of Z are referred to as Freeman-Tukey deviates.

FreemanTukey transformation formula for Poisson variables



- By applying the Freeman-Tukey transformation, you can achieve normality of the data, which is a requirement for many statistical tests. This allows you to perform reliable statistical analysis on proportions and Poisson variables, even when the data comes from small samples or has extreme values.

Square Root and Arcsine

- The square root and arcsine (inverse sine) functions are both mathematical operations, but they play a specific role in the Freeman-Tukey transformation.

Square Root

- In the Freeman-Tukey transformation for proportions, the square root is taken first. This helps to compress the range of values, particularly for proportions close to 0 or 1. Since the square root of a small number is also small, it reduces the influence of extreme values on the transformation.

Arcsine

- After taking the square root, the arcsine function is applied. The arcsine function takes an input between -1 and 1 and returns the angle in radians whose sine is that value. This step converts the compressed values from the square root operation into a range suitable for a normal distribution.

Combined Effect

By combining the square root and arcsine, the Freeman-Tukey transformation achieves two things:

1. **Reduces the influence of extreme values:** The square root helps compress the range, making proportions closer to 0 or 1 less impactful on the overall distribution.
2. **Transforms the data towards normality:** The arcsine function converts the compressed values into a form that more closely resembles a normal distribution.

Log and Exponential transforms

- Log and exponential transforms are another set of powerful tools used in data analysis, distinct from the Freeman-Tukey transformation.

Logarithmic Transformations

- **Purpose:** Logarithms are used to compress data that exhibits exponential growth or to address skewed distributions. They work by taking a value (x) and converting it to its exponent in relation to a base (a), written as log_a(x). The most common base used in statistics is the natural base (e), resulting in the natural logarithm (ln(x)).

Applications:

- **Compressing skewed data:** When dealing with data that grows exponentially or has a right skew (concentrated on the left side), logarithms can compress the larger values, making the distribution more symmetrical.
- **Analyzing ratios and changes:** Logarithms are helpful for analyzing ratios and percentage changes. For example, comparing the log income of different groups can reveal trends in income inequality.

Exponential Transformations

- **Purpose:** Exponential transformations reverse the effect of logarithms and are used to model exponential growth or decay. They raise a base (a) to the power of another value (x), written as a^x.

- **Applications: Modeling exponential growth:** Exponential functions are ideal for modeling phenomena with rapid growth, such as population increase, bacterial growth, or radioactive decay.
- **Creating interaction terms:** In regression analysis, exponential terms can be used to capture interactions between variables where the effect of one variable on another is not constant.

Choosing the Right Transform

The choice between using a log or an exponential transform depends on the characteristics of your data and the analysis you want to perform.

- Use logarithms if your data exhibits exponential growth or has a right skew.
- Use exponentials if you want to model exponential relationships or create interaction terms in regression analysis.

In Relation to Freeman-Tukey

- The Freeman-Tukey transformation is specifically designed for normalizing proportions, while log and exponential transforms address broader data issues like skewness and exponential relationships. They can be complementary tools:
- You might use the Freeman-Tukey transformation on proportions before applying a log transform to further address skewness.
- Conversely, you could use a log transform on a variable before incorporating it into a model where you suspect an exponential relationship.

Logit transforms

- The logit transformation occupies a unique space compared to the previously discussed transformations (Freeman-Tukey, log, and exponential).
- **Purpose**: The logit transform, also known as the logistic transformation, is primarily used to convert binary data (0 or 1) or proportions (between 0 and 1) into a continuous value between negative infinity and positive infinity.
- **Formula**: The logit of a proportion (p) is calculated using the natural logarithm (ln) of the odds: $\ln(p / (1 - p))$.

- **Applications:**
- **Logistic Regression:** The logit transformation is the foundation for logistic regression, a statistical method used to model the relationship between a binary dependent variable (e.g., success/failure, alive/dead) and one or more independent variables. The transformed values (logits) allow logistic regression to estimate the probability of an event occurring based on the independent variables.
- **Visualizing Binary Data:** Logit transformed data can be plotted on a continuous scale, making it easier to visualize the relationship between the independent variable and the probability of the event.

Differences from Other Transforms

- **Focus on Binary Data**: Unlike log and exponential transforms that address a wider range of data characteristics, the logit focuses specifically on transforming binary data or proportions.
- **Normalization for Logistic Regression**: The logit transformation doesn't necessarily create a normal distribution, but it transforms the data into a scale suitable for logistic regression analysis.
- **Odds and Probability**: The logit transformation works with the concept of odds (p / (1 - p)) instead of directly manipulating the data values themselves.

Relationship with Other Transforms

- **Complementary to Log Transformation**: In some cases, the logit transformation can be seen as an alternative to the log transform for proportions, especially when the focus is on modeling the probability of an event using logistic regression.
- **Not Directly Related to Freeman-Tukey**: The Freeman-Tukey transformation aims for general normality of proportions, while the logit serves a specific purpose in logistic regression. However, both can be used on proportions depending on the analysis goals.

Normal transformation

- The standard normal transformation, also known as the z-score transformation, is a specific technique used to convert data from any distribution into a standard normal distribution. A standard normal distribution has a mean of 0 and a standard deviation of 1. This transformation allows you to compare data points from originally different distributions using a common scale.

Formula:

- The transformation is achieved using the following formula:
- $z = (x - \mu) / \sigma$

Where:

- * z is the transformed value (z-score)
- * x is the original data value
- * μ is the mean of the original data distribution
- * σ is the standard deviation of the original data distribution

- **Effect:** This formula subtracts the mean (μ) from each data point (x) and then divides by the standard deviation (σ). This process effectively centers the data around 0 and scales it to have a standard deviation of 1.

Benefits of Standard Normal Transformation

- **Standardization:** By transforming data into a standard normal distribution, you can compare values from originally different distributions on a common scale (z-scores). This allows you to assess the relative position of a data point within its original distribution.
- **Statistical Tests:** Many statistical tests rely on the assumption of normality. The standard normal transformation can be a helpful step before applying these tests to non-normal data. By transforming the data to a normal distribution, you can ensure the validity of the test results.