Data Screening, Diagnostics, and Transformations

Steven M. Boker

Department of Psychology
University of Virginia

Structural Equation Modeling
Psyc–8501–001

Overview

- Plotting Univariate Distributions
- Histograms.
- Box Plots.
- QQ Plots.
- Matrix Scatter Plots.

Some Definitions

- *Order Statistics* are based on ordering (sorting or ranking) the data.
- Order Statistics tend to be robust in that they are resistant to the influence of outliers.
- *Quantiles* are selected ranges of the ordered data based on percentages of the total N.
- The *Depth* of an element is the number of elements in the ordered set that are between the chosen element and the nearest end of the set.

Order Statistics

- The Data

  1 3 6 8 9 6 2 1 5

- Sort the data

  1 1 2 3 5 6 6 8 9

- Depth

  1 2 3 4 5 4 3 2 1

- Order Statistics: *Extreme*, *Hinge*, and *Median*

  E H M H E

  1 1 2 3 5 6 6 8 9

Extremes and Median

- The *Extremes* of a dataset are the values of the elements whose depth is 1.
- The *Median* of a dataset is the value of element whose depth is $(N+1)/2$.
  - If N is odd, then the median takes on the value of the element whose depth is $(N+1)/2$.
  - If N is even, then the median takes on the average of the values of the two elements whose depths are N/2.

Hinges, Interquartile Range, and Adjacent Values

- The *Hinges* are the values of the elements whose depth are midway between the median and the extremes.

$$d(H) = \frac{d(M) + 1}{2}$$

- Apply the same rules for even and odd depths as were used for the median.
- The *Interquartile Range* is the difference between the values of the two hinges.
- *Adjacent Values* are the values that are equal to the upper (or lower) hinge plus (or minus) 1.5 times the interquartile range.

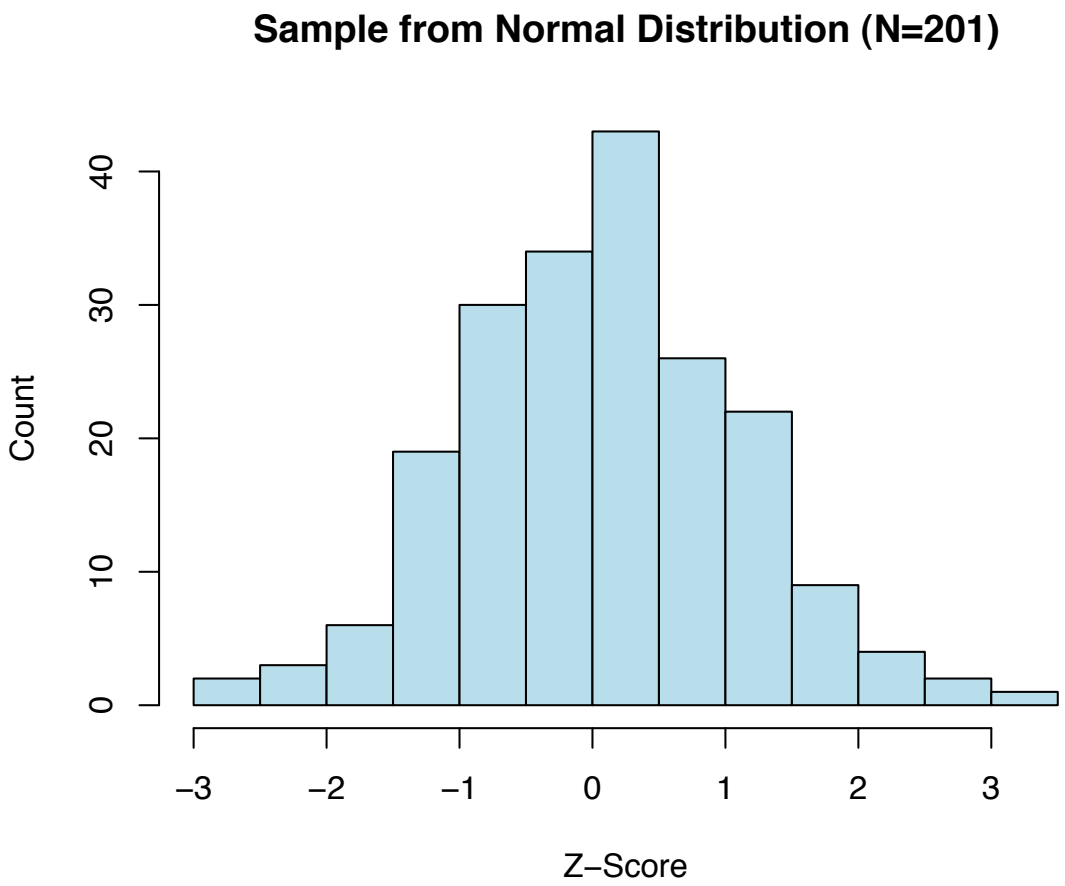
Boxplots

- A *Boxplot* (or Box-and-Whisker Plot) plots the median, hinges, interquartile ranges, adjacent values and all observations more extreme than the adjacent values.
- In a boxplot, alternative names for the adjacent values are whiskers or fences.

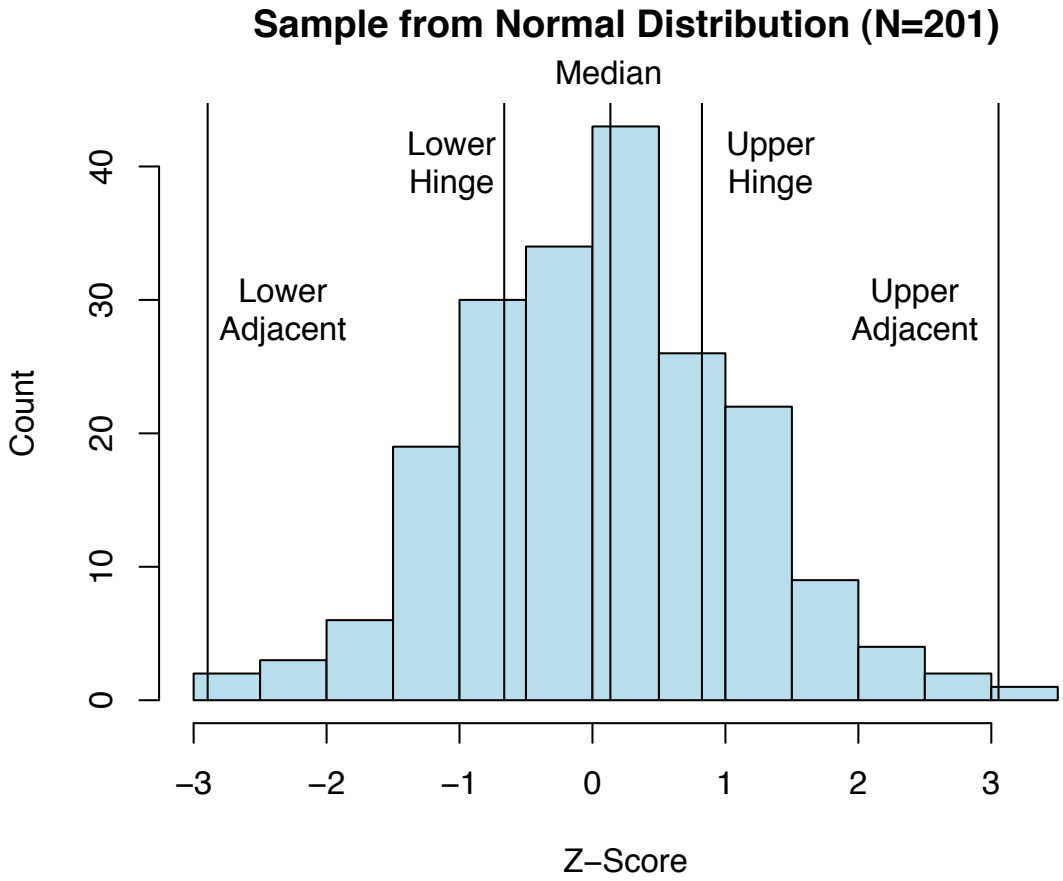
Box Plot Example

- Draw a sample of $N = 201$ from a normal distribution with mean of 0 and standard deviation of 1.
- Create a histogram of the sample.
- Create a boxplot of the sample.

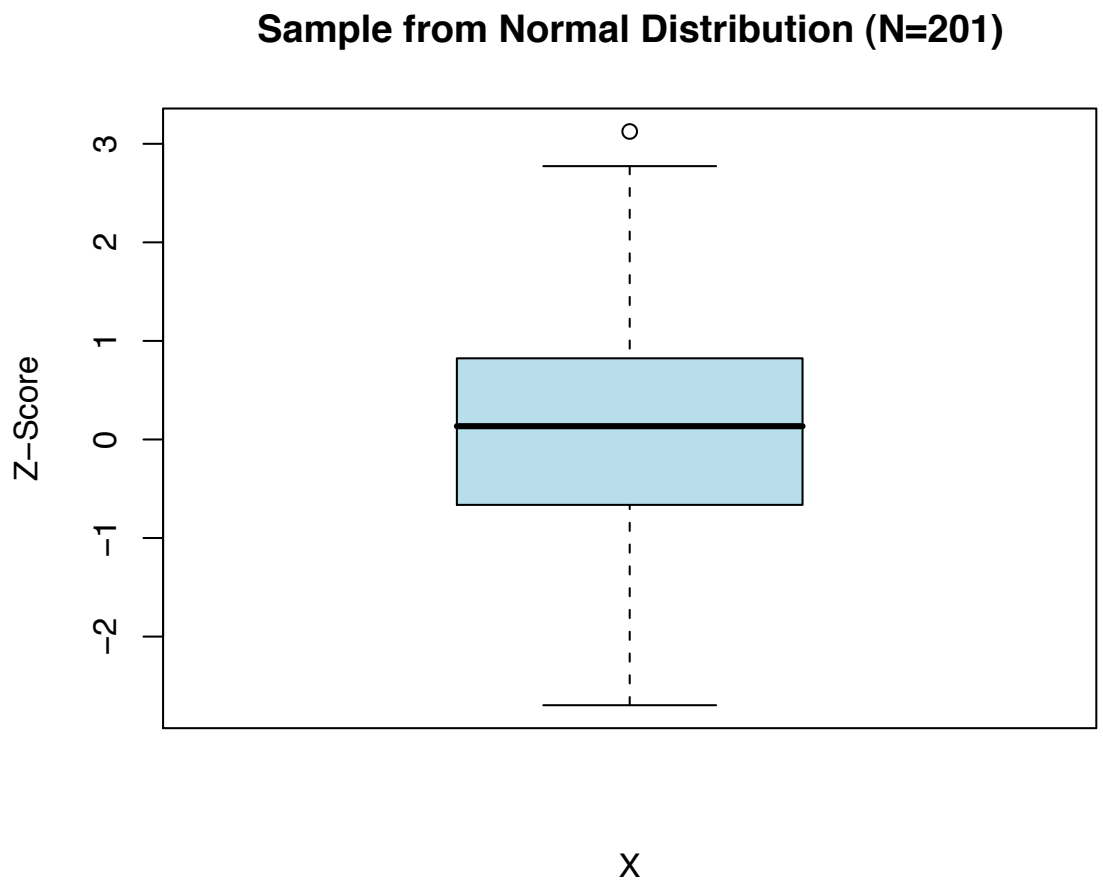
Histogram Example



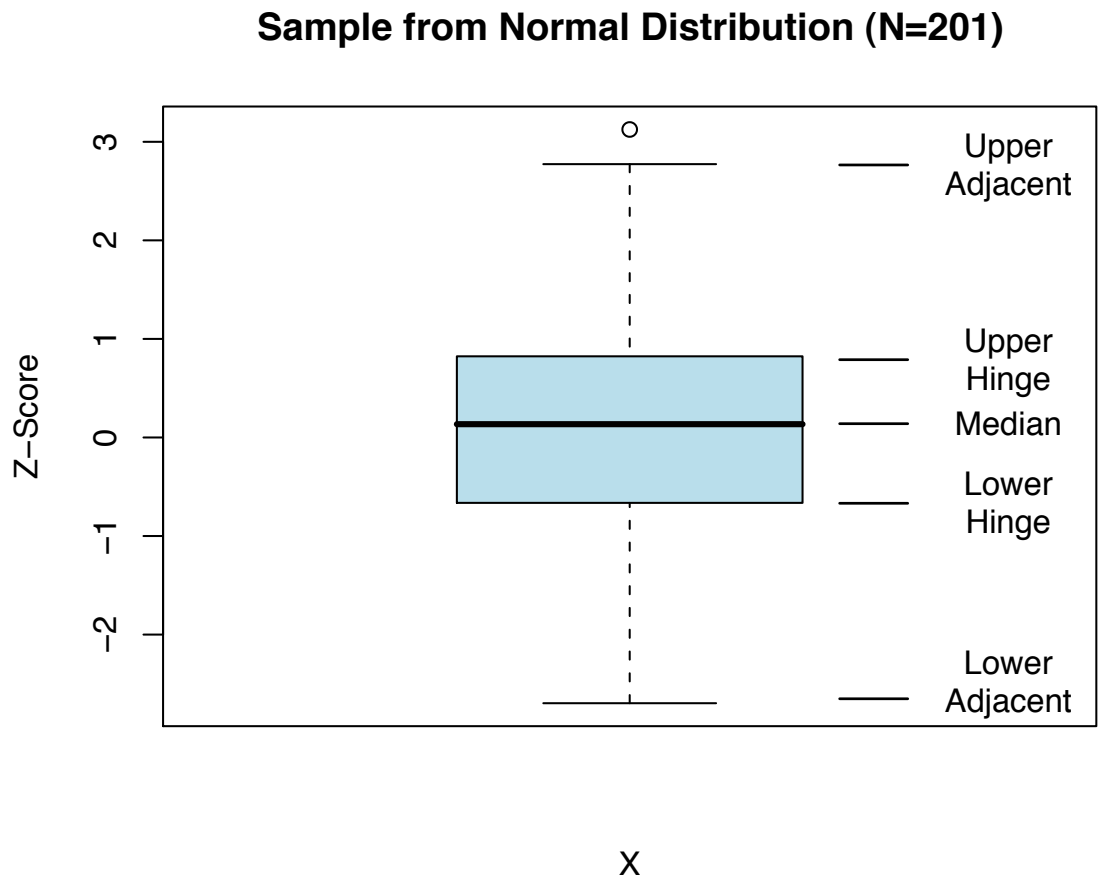
Histogram Example



Box Plot Example



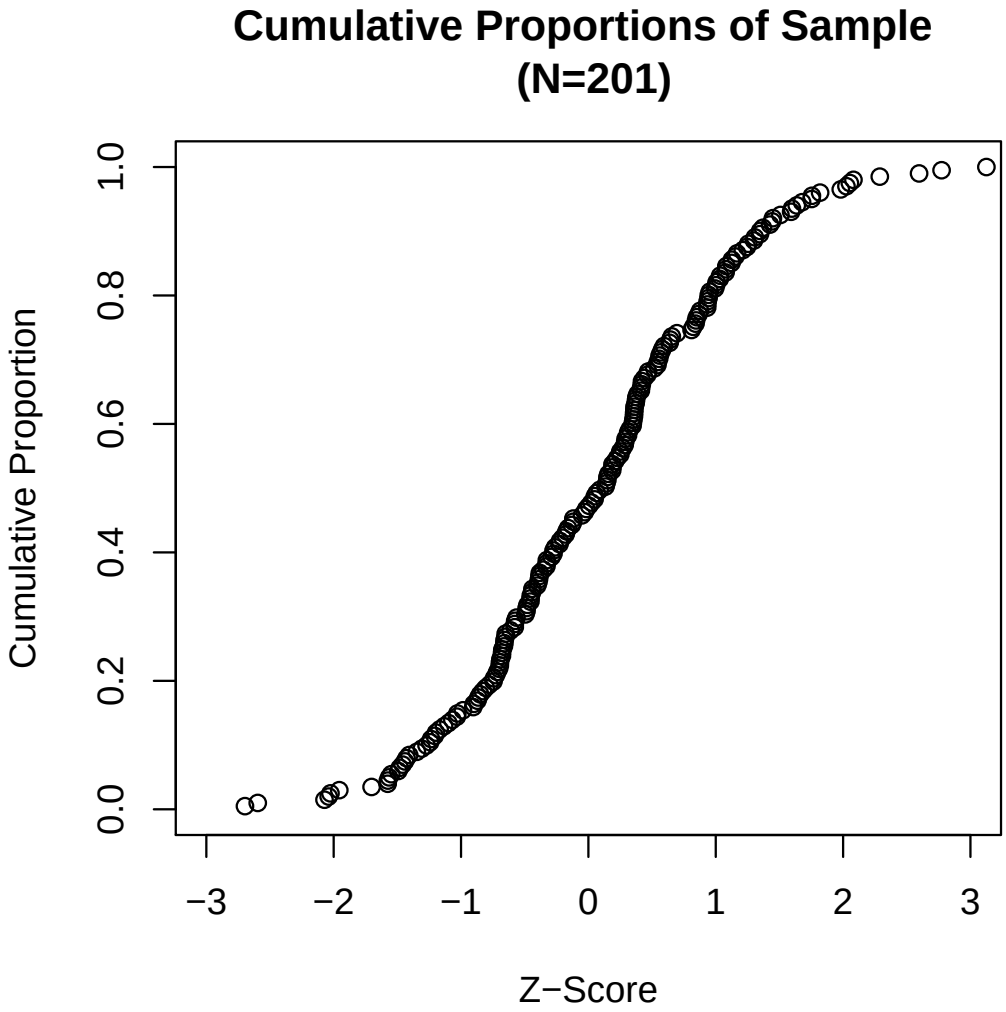
Box Plot Example



Cumulative Proportion Plot

- A Cumulative Proportion Plot displays the cumulative proportion of all values less than or equal to each value in the dataset.
- This essentially plots a variable's quantiles against its values.

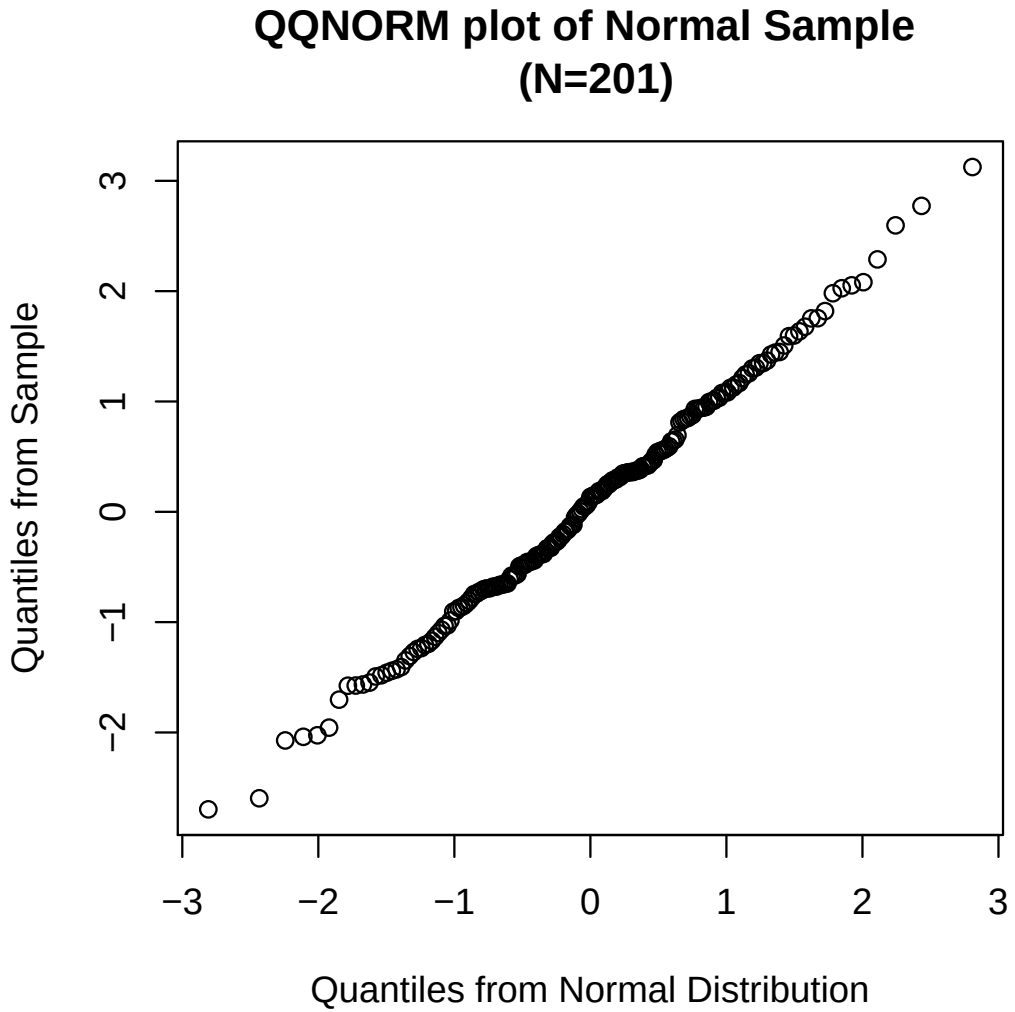
Cumulative Proportion Example



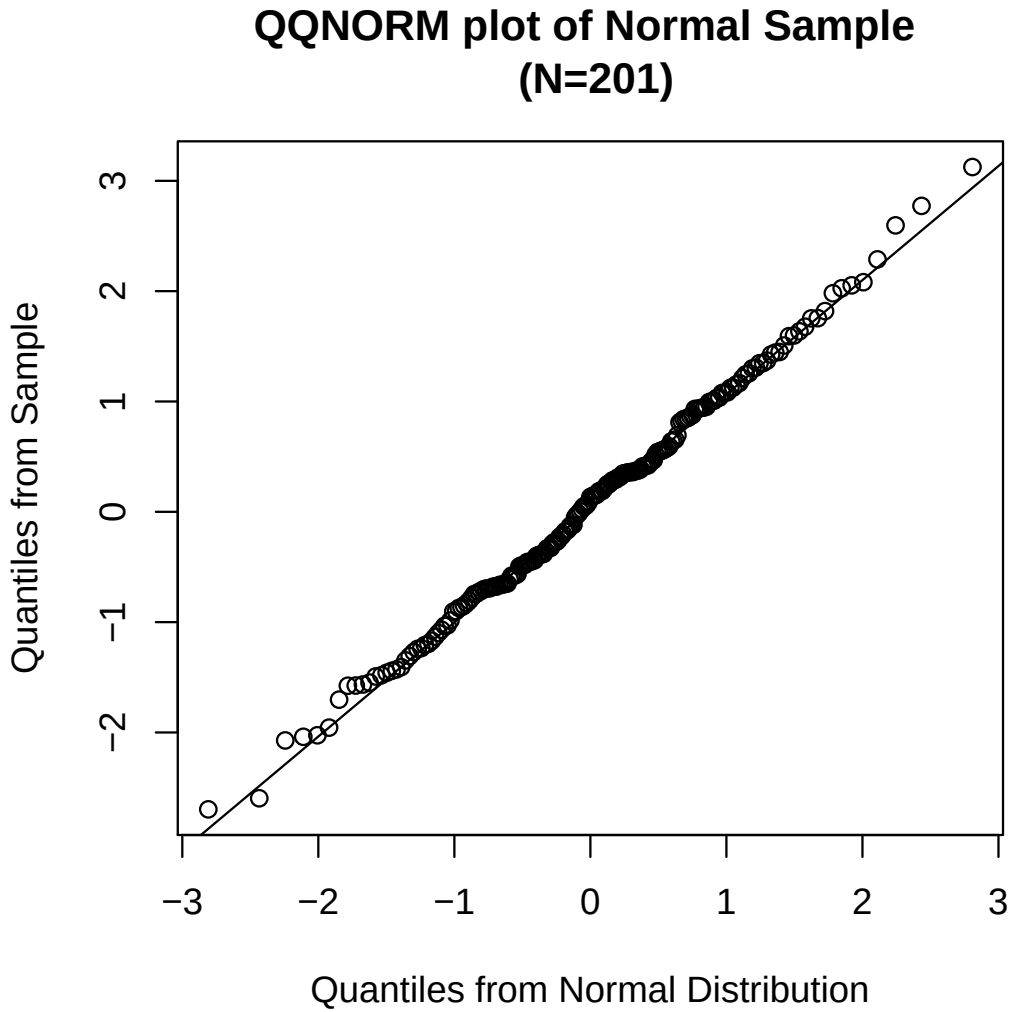
Quantile–Quantile (QQ) Plots

- Quantile–Quantile Plots compare two distributions.
- Quantile–Normal Plots compare the data against a normal distribution.

QQ Normal Example



QQ Normal Example with Regression Line



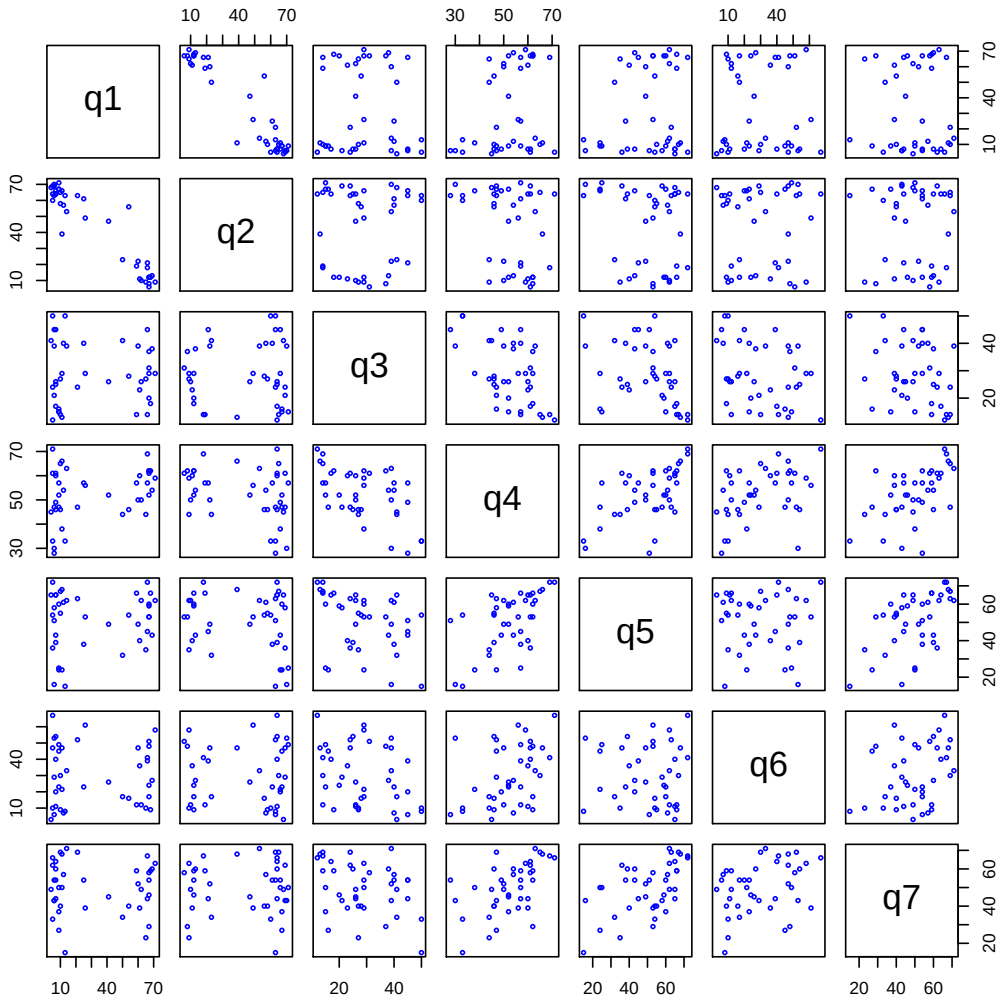
Interpreting a Quantile Quantile Plot

- Difference between the means (medians)
  - The midpoint of the plotted points is off the diagonal.
- Difference between the variances
  - The plotted points fit a line that is not parallel to the diagonal.
- Mixture distributions
  - Some parts of the distribution fit one line (or curve) and other parts fit another line (or curve).
- Different type of distributions entirely
  - The plotted points do not fit a line.

Matrix Scatterplot

- A matrix scatterplot, or a *pairs plot* in R, plots bivariate relationships between multiple variables.
- You can plot either an R matrix or dataframe.
- Pairs plots are among the first plots I make when I see a new dataset.
- This can guide you very quickly towards what is happening in your data.

Matrix Scatterplot



Outline

- Why Transformations?
- Common Transformations.
  - Log transform
  - Power transforms
- Centered Power Transform

Advantages of Transforming Variables

- Make a more meaningful scale.
- Conform to assumptions of normality for linear models.
- Create comparable scales
  - Useful in psychometric measurement.
- Learn about possible processes that might have generated an observed distribution.

Disadvantages of Transforming Variables

- Might lose the meaning of a scale.
- Might lose scale comparability.
- Could be difficult to explain scale to readers.
- Reviewers and readers may consider transformations to be arbitrary or difficult to understand.

The Two Most Commonly Used Psychological Transformations of All Time

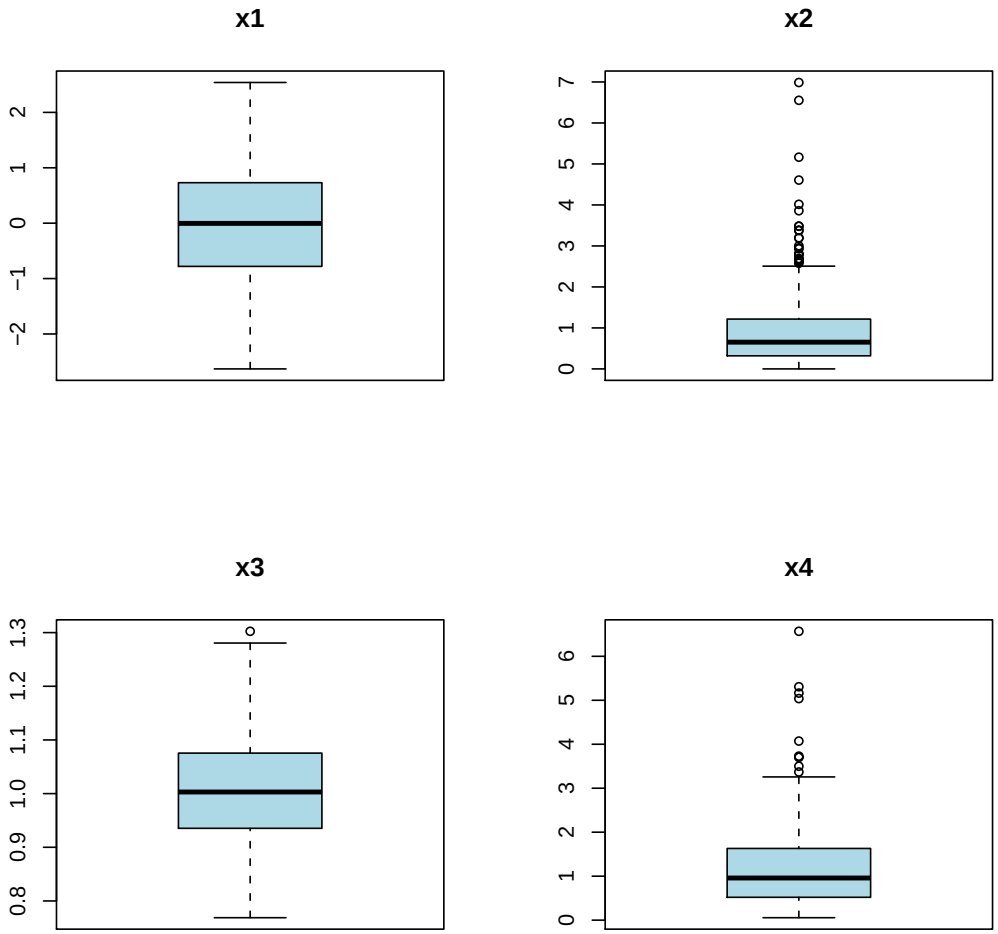
- Centering (subtracting the mean)
  - Does not alter the shape of the distribution.
  - Does not change the scale.
- The Z–Score Transformation
  - Does not alter the shape of the distribution.
  - Does change the scale.

Four Example Distributions

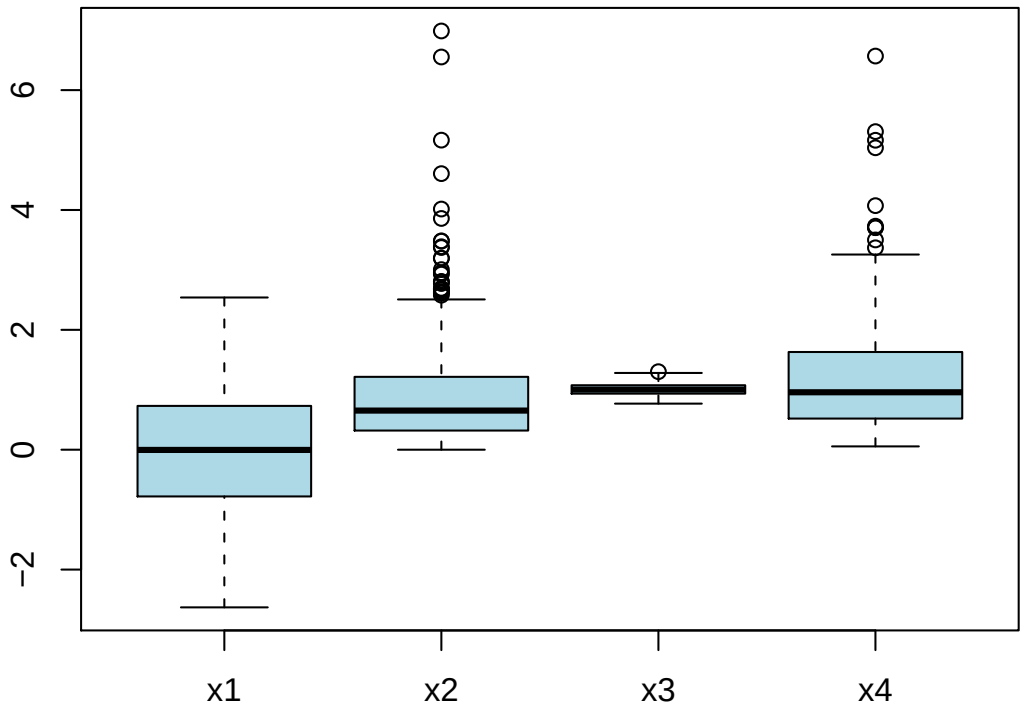
- Here are four example distributions.

```
x1 <- rnorm(300, mean=0, sd=1)
x2 <- rexp(300)
x3 <- (1 / rnorm(300, mean=1, sd=.1))
x4 <- rnorm(300, mean=1, sd=.2)^4
```

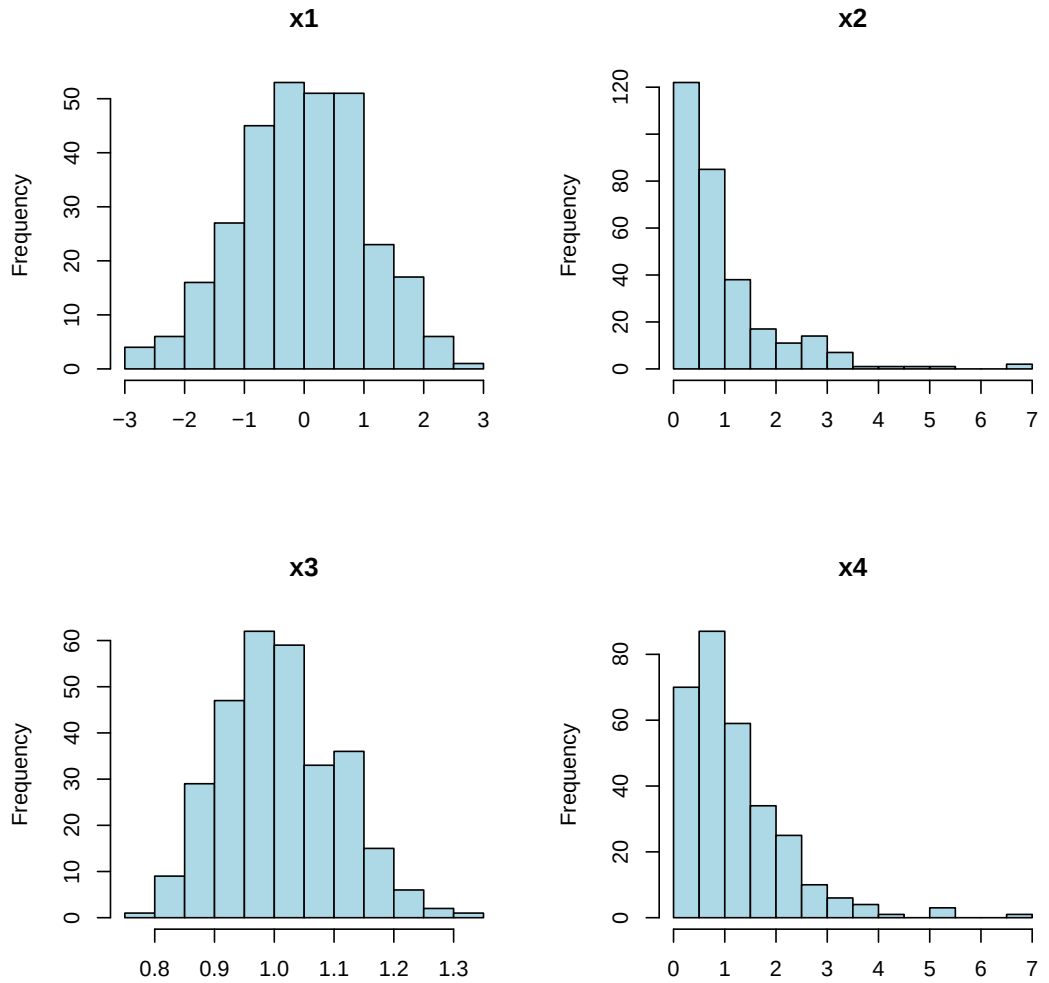
Four Example Distributions, Separate Scales



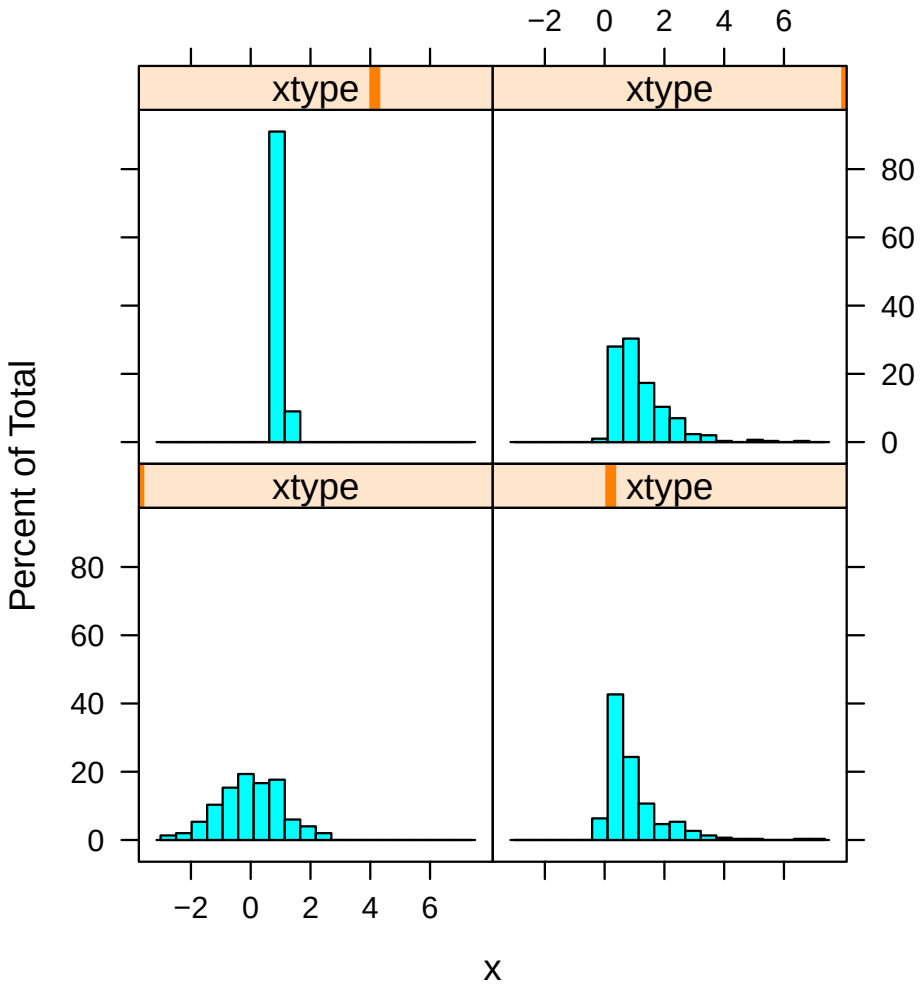
Four Example Distributions, Same Scale



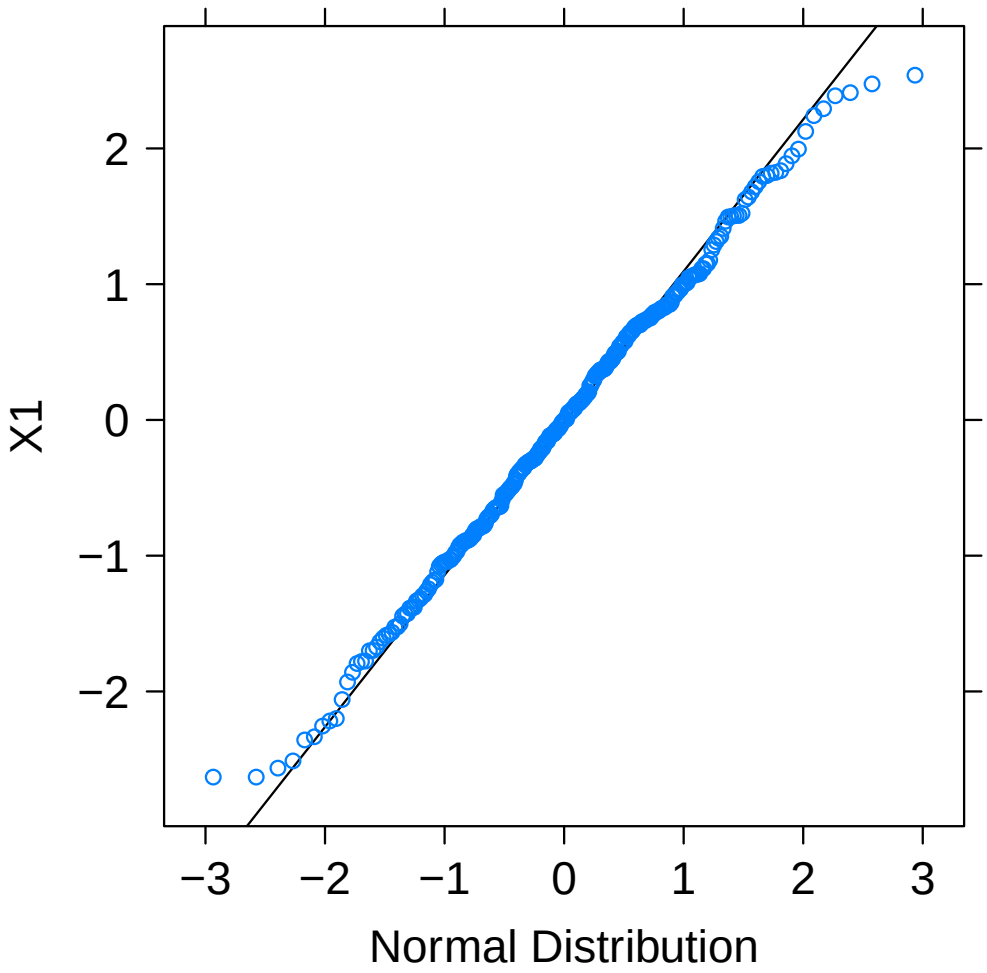
Four Example Distributions, Separate Scales



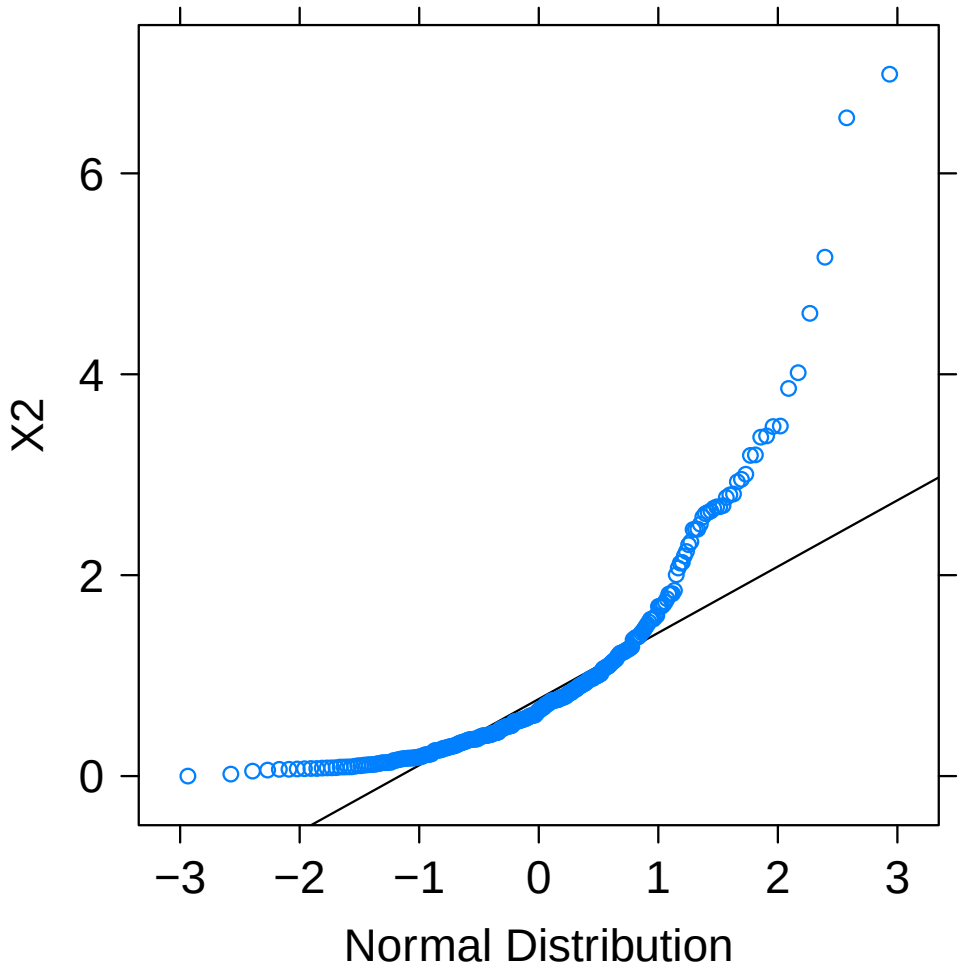
Four Example Distributions, Same Scale



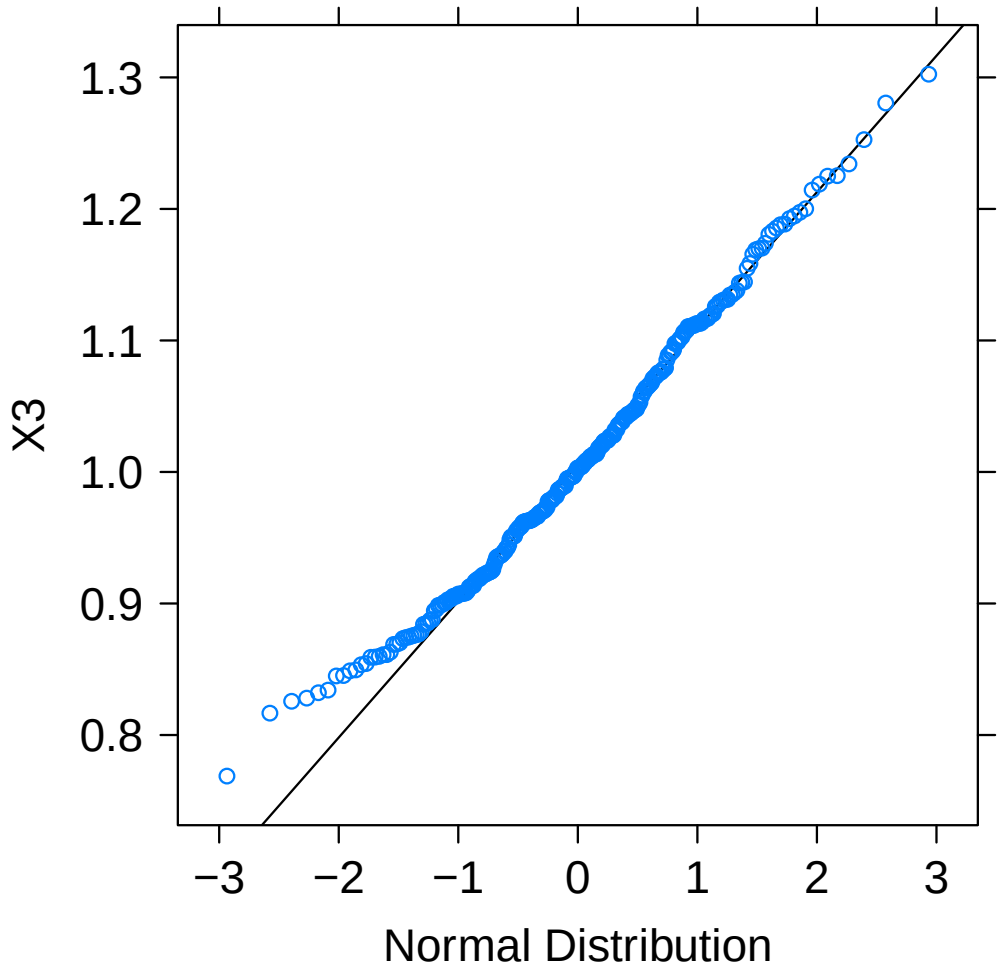
Normal Distribution



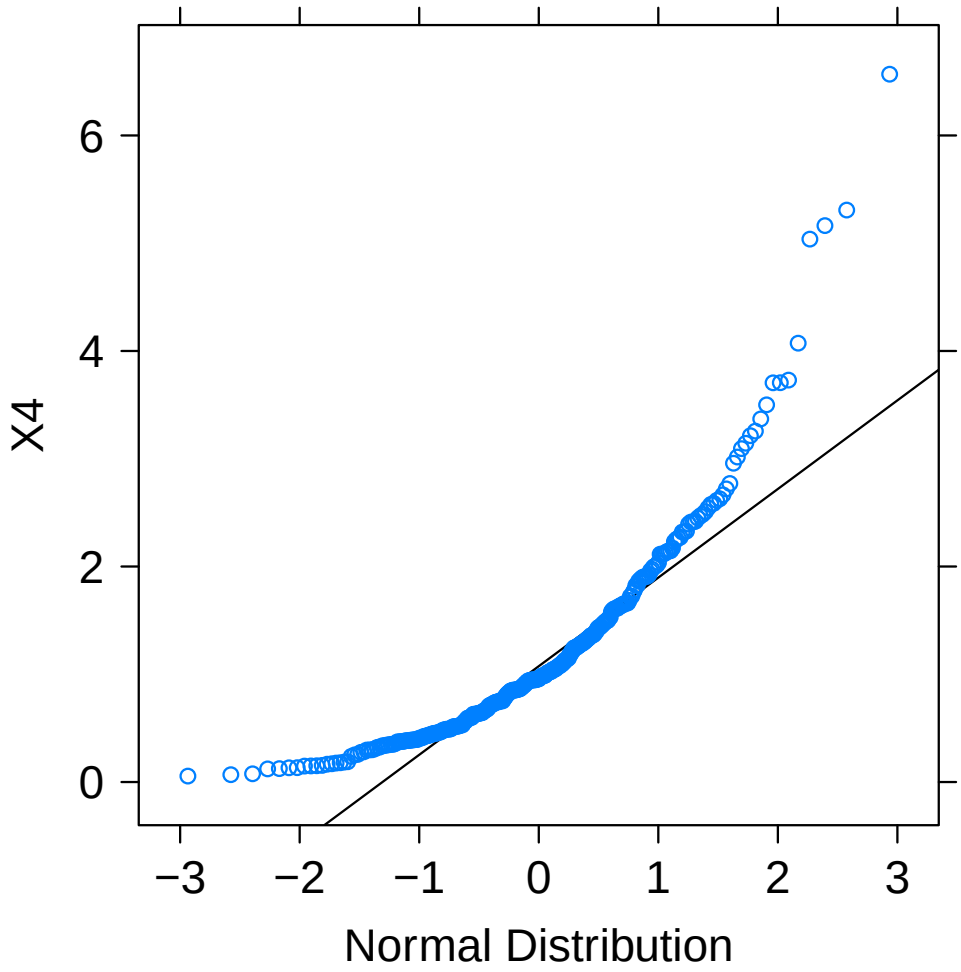
Exponential Distribution



Inverse Normal Distribution with Mean=1



Normal Distribution to the Fourth Power



Log Transformations

- Natural log of the vector **x4** can be calculated as:

  `log(x4)`

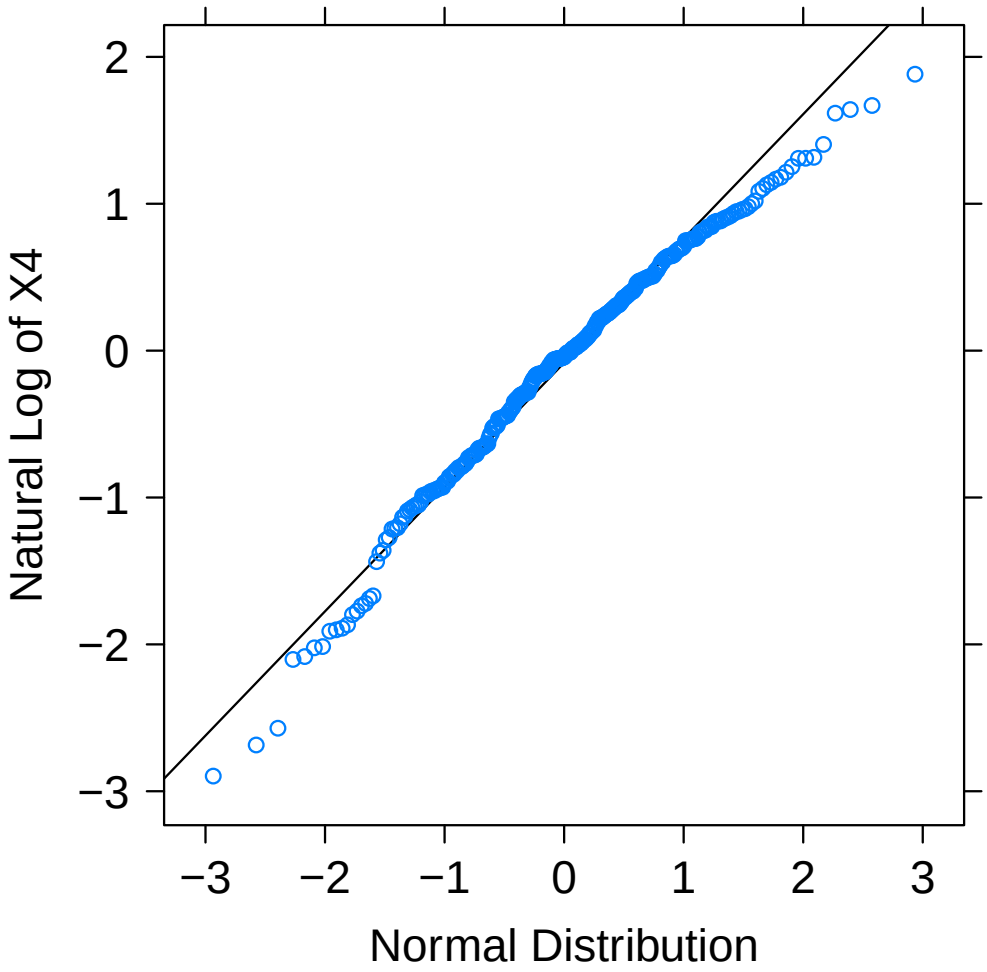
- Log base 10 and log base 2 are:

  `log10(x4)`

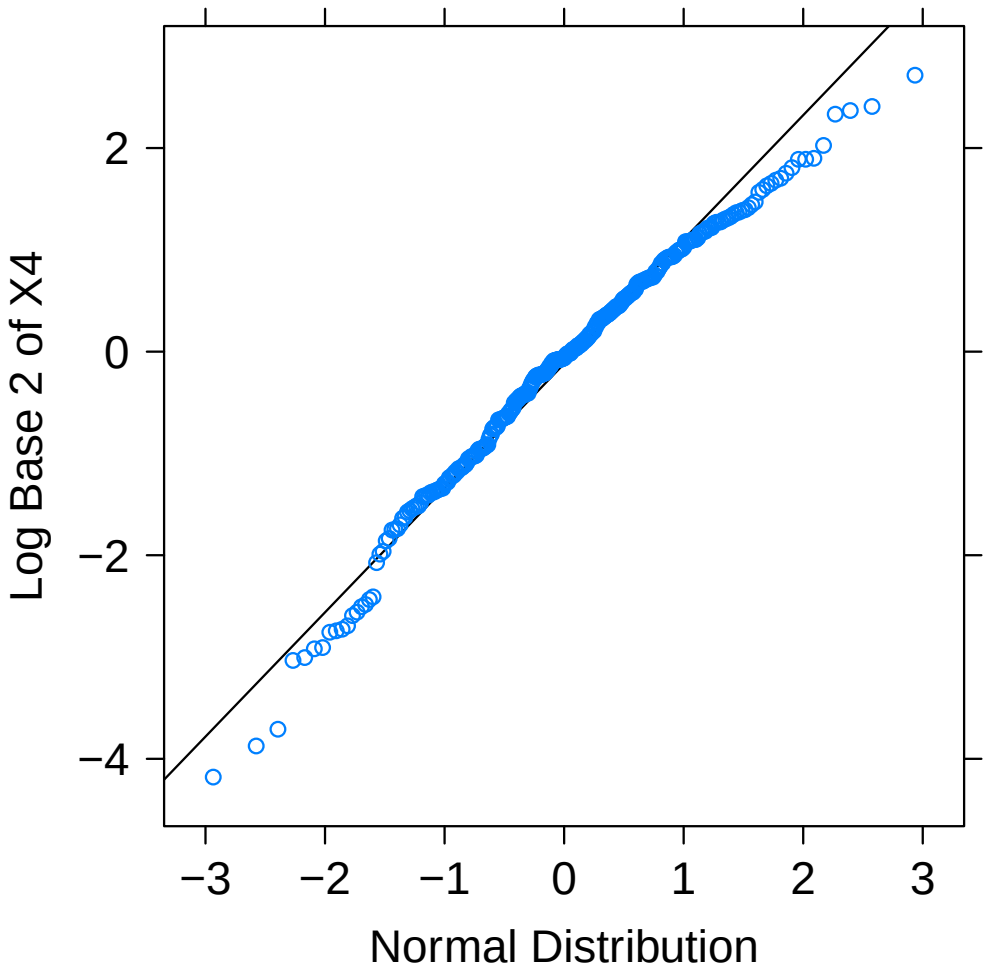
  `log2(x4)`

- Does it matter which base you choose?

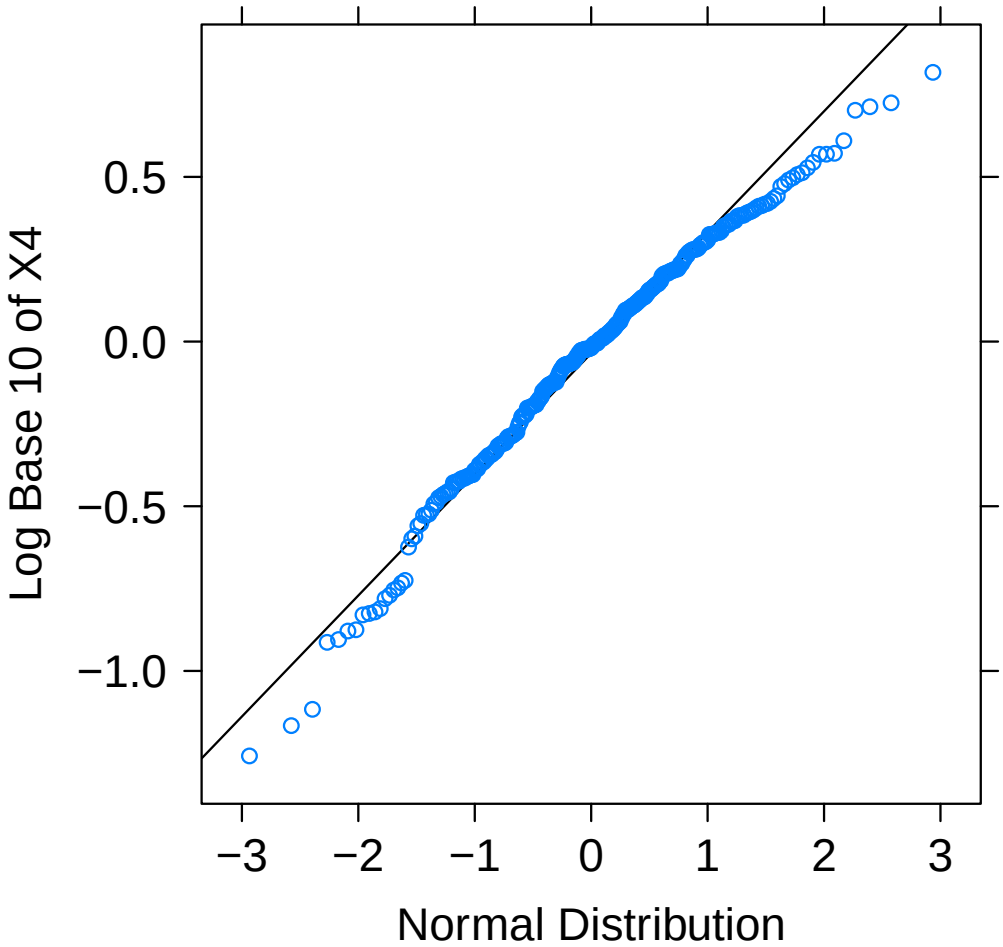
Natural Log Transformation



Log Base 2 Transformation



Log Base 10 Transformation



Interpreting Log Transformed Scales

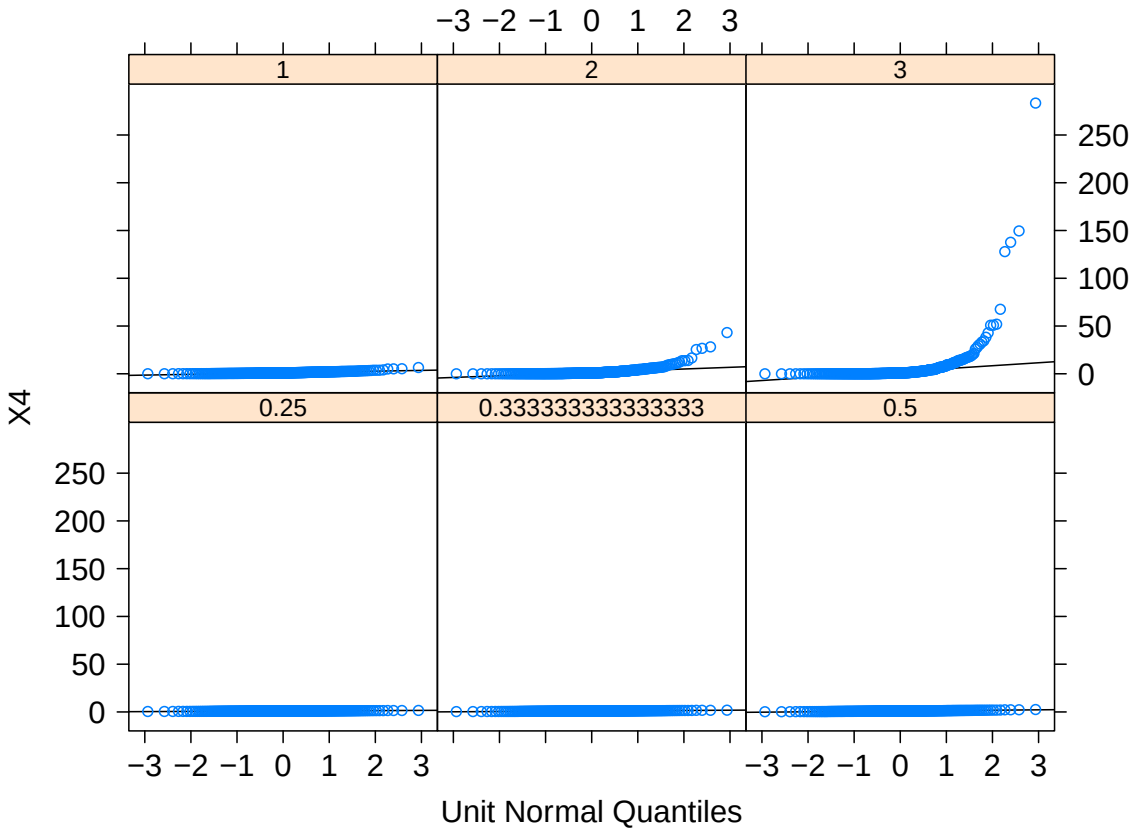
- Log transform changes the scale (from say seconds to log-seconds).
- Addition between log units is like multiplication between regular units.
- Thus the difference between two means in log-units represents how many times greater the larger mean is than the smaller mean in regular units.

Power Functions

- We can try a variety of power functions using QQ plots.
- Create a splitting variable and a transformed vector.

```
tlen <- length(x4)
exponent <- c(rep(1/4, tlen),
              rep(1/3, tlen),
              rep(1/2, tlen),
              rep(1, tlen),
              rep(2, tlen),
              rep(3, tlen))
tempx2 <- c(x4^(1/4), x4^(1/3), x4^(1/2),
            x4^1, x4^2, x4^3)
```

Power Transformations

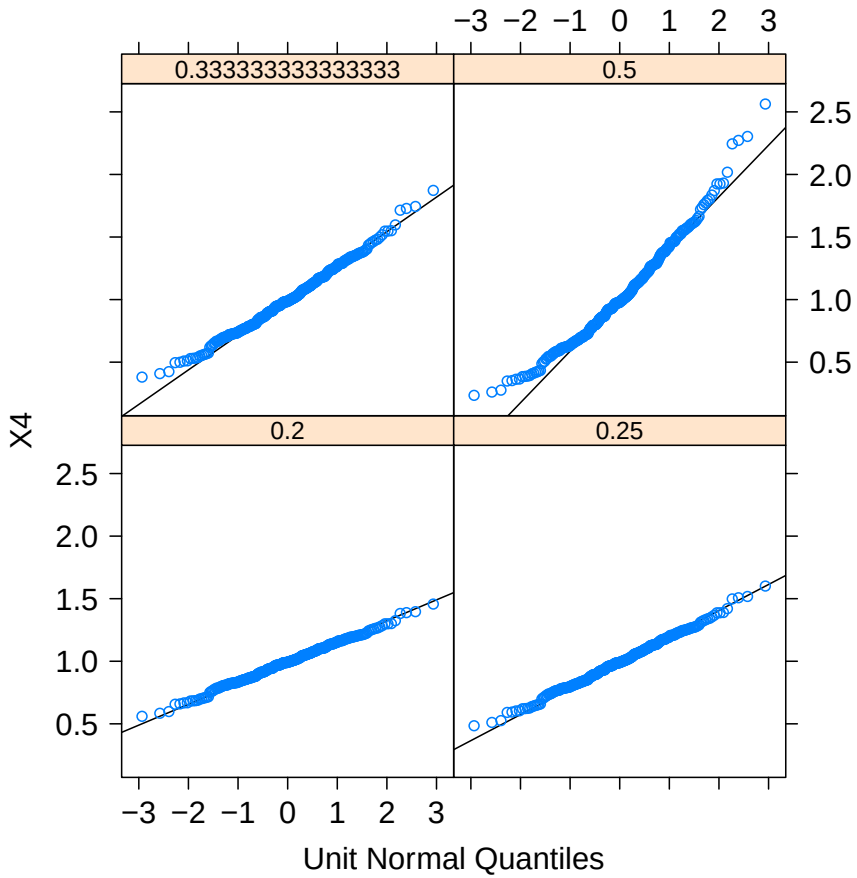


Power Functions

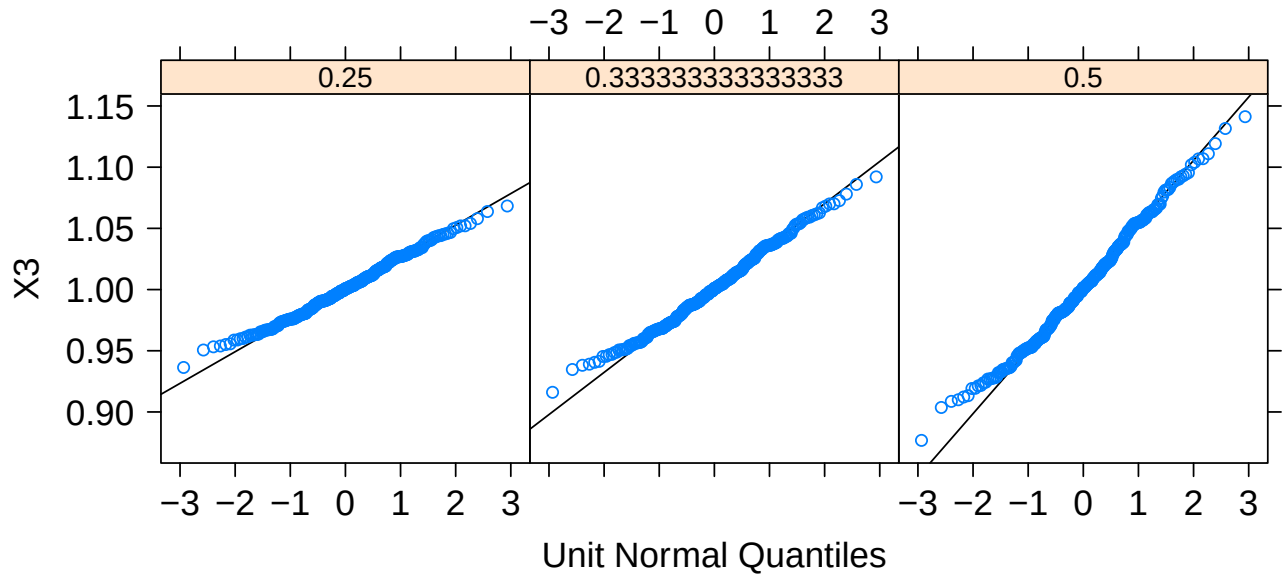
- Let's just try the fractional exponents.

```
tlen <- length(x4)
exponent <- c(rep(1/5, tlen),
              rep(1/4, tlen),
              rep(1/3, tlen),
              rep(1/2, tlen))
tempx2 <- c(x4^(1/5), x4^(1/4),
            x4^(1/3), x4^(1/2))
```

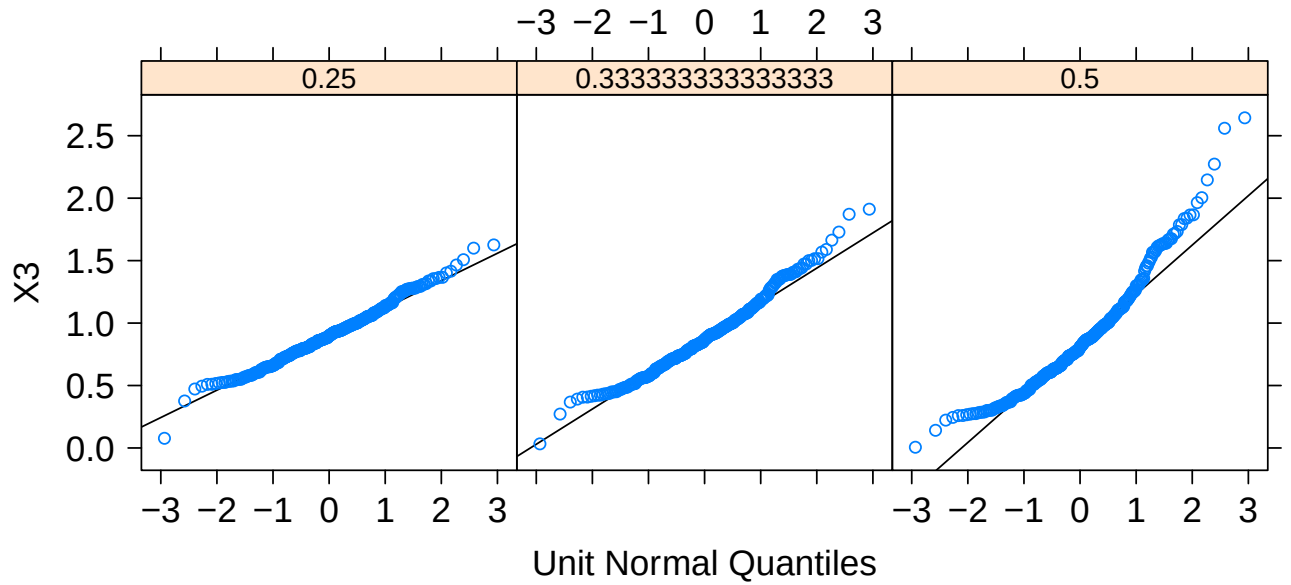
Power Transformations



Power Transformation on an Inverse Normal



Power Transformation on an Exponential Distribution

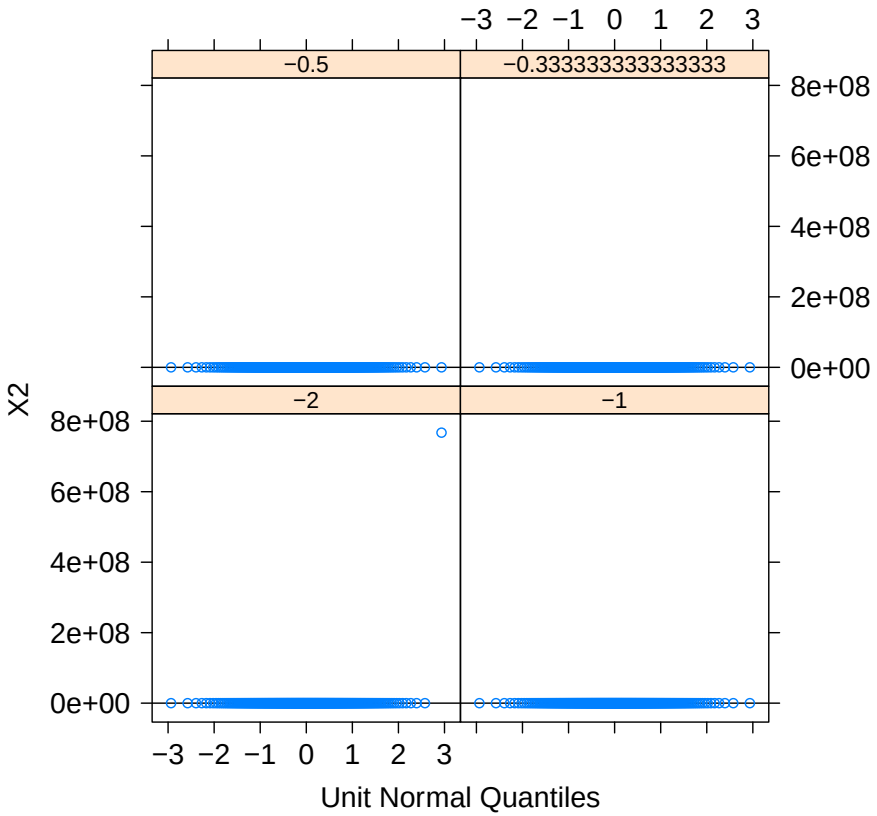


Negative Exponent Power Functions

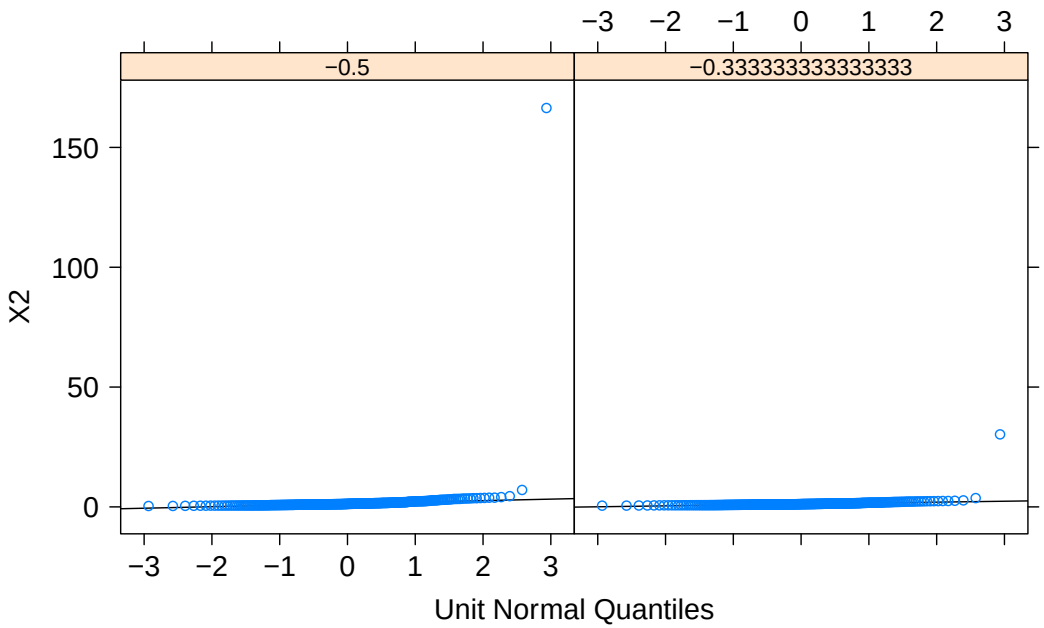
- How about negative exponents?

```
tlen <- length(x3)
exponent <- c(rep(-1/3, tlen),
              rep(-1/2, tlen),
              rep(-1, tlen),
              rep(-2, tlen))
tempx2 <- c(x3^(-1/3), x3^(-1/2),
            x3^-1, x3^-2)
```

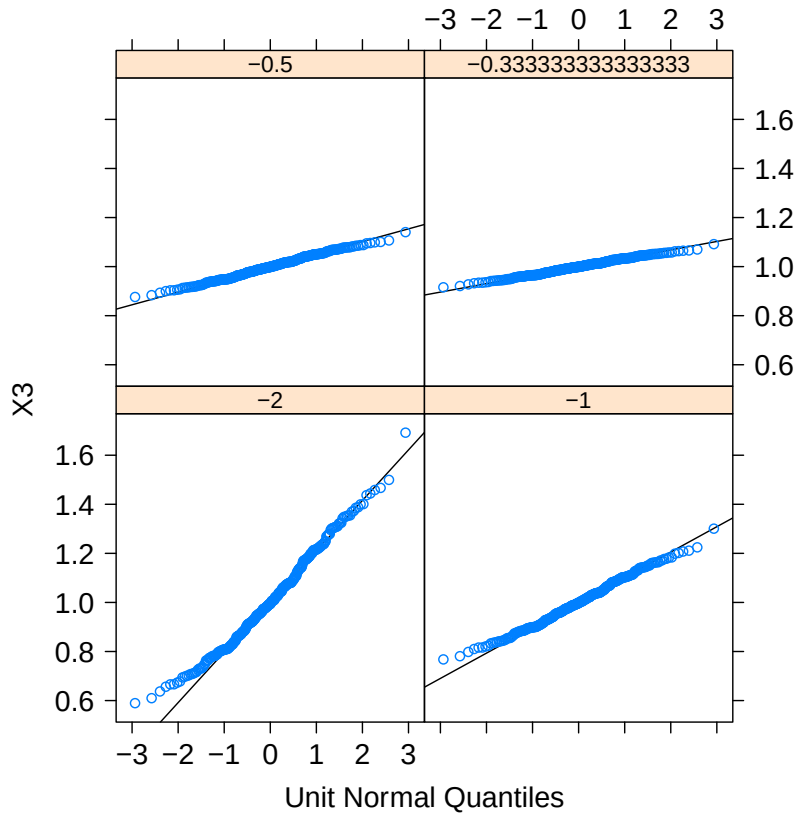
Negative Exponent Power on an Exponential



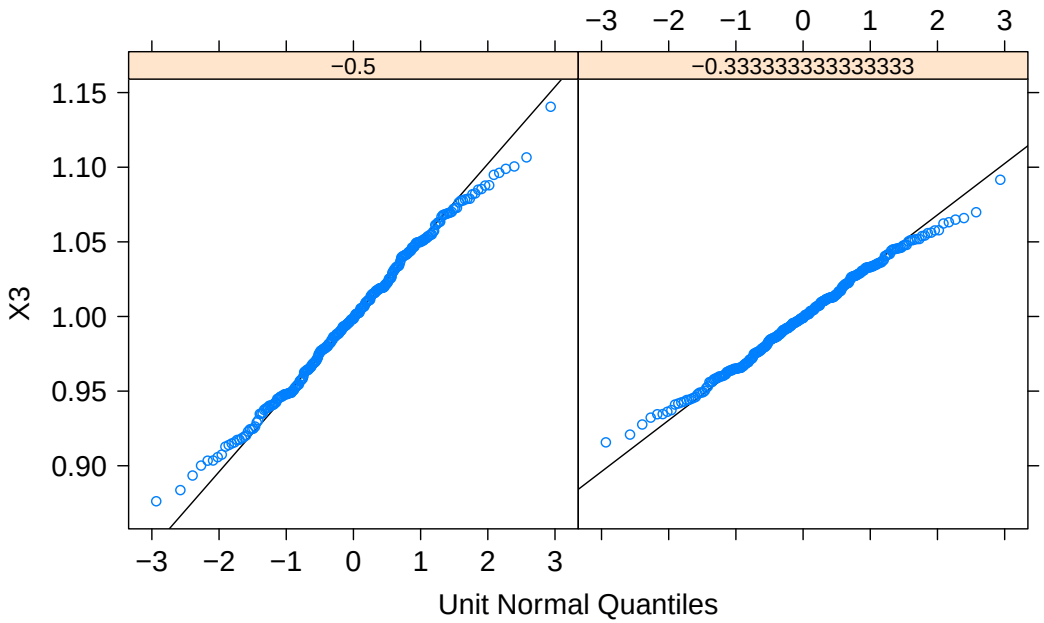
Negative Exponent Power on an Exponential



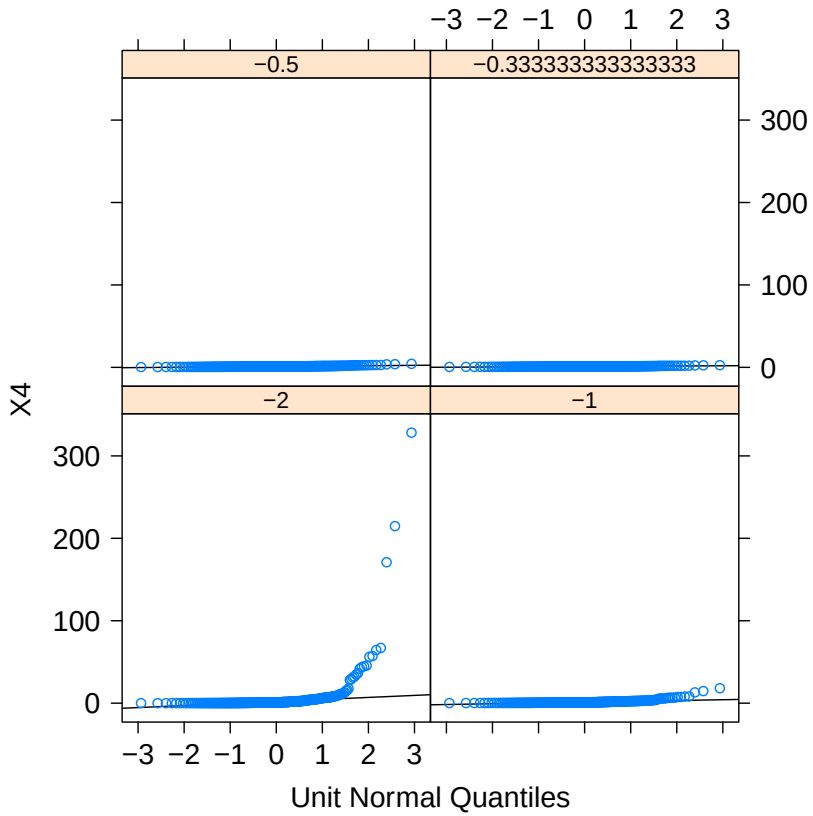
Negative Exponent Power on an Inverse Normal



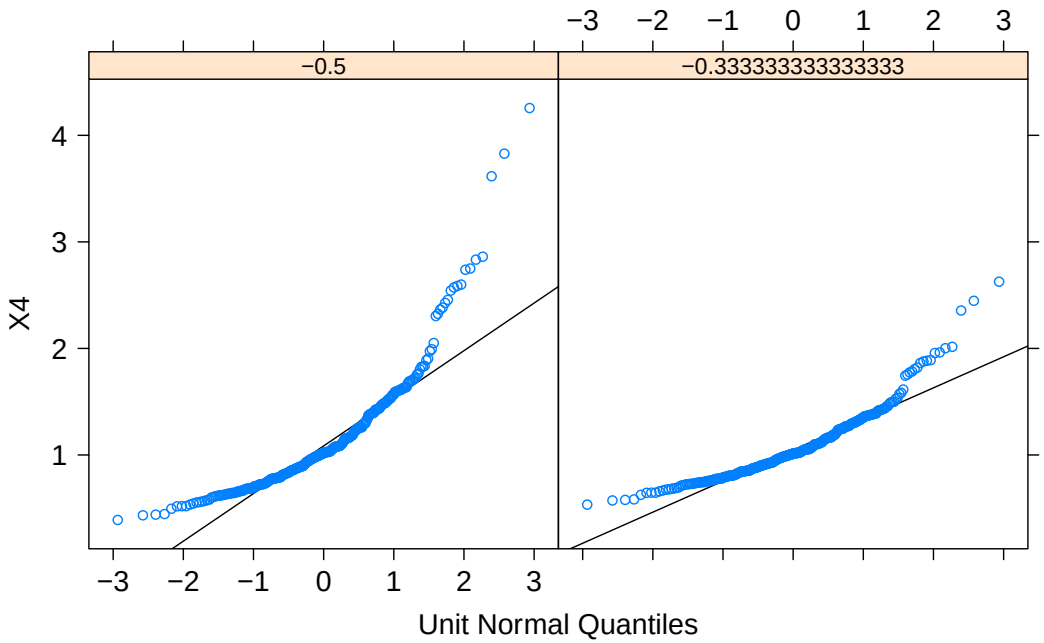
Negative Exponent Power on an Inverse Normal



Negative Exponent Power on a Positive Exponent



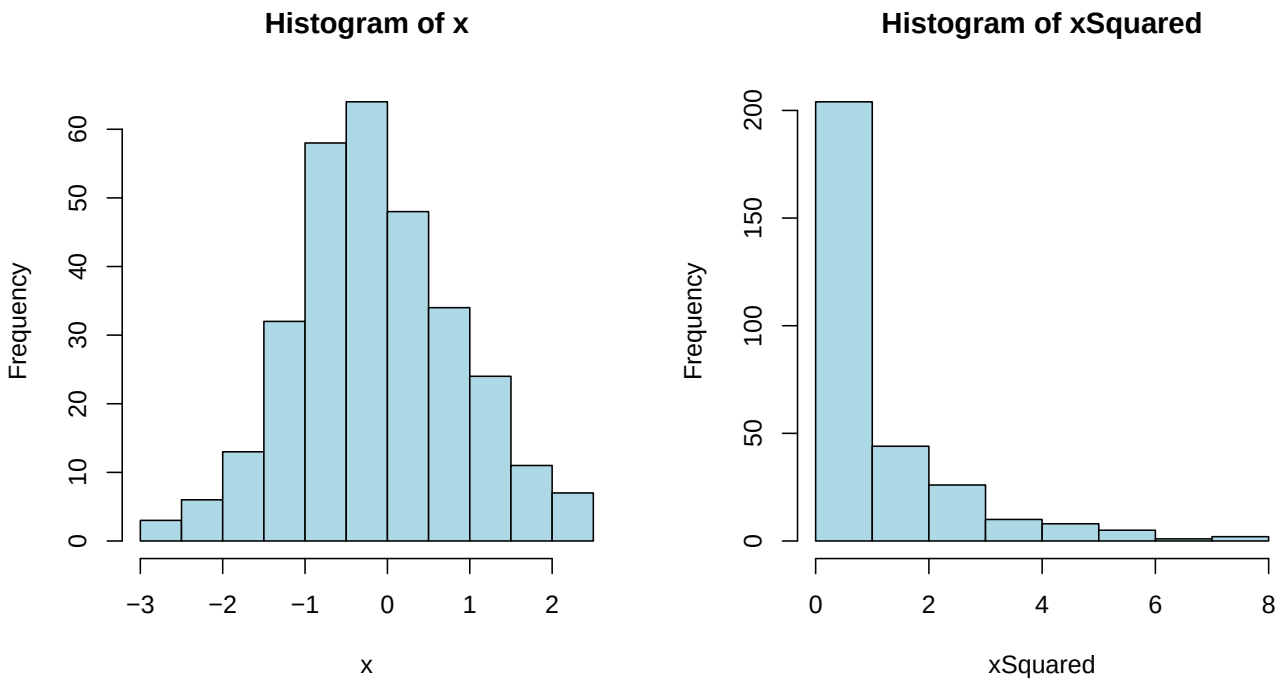
Negative Exponent Power on a Positive Exponent



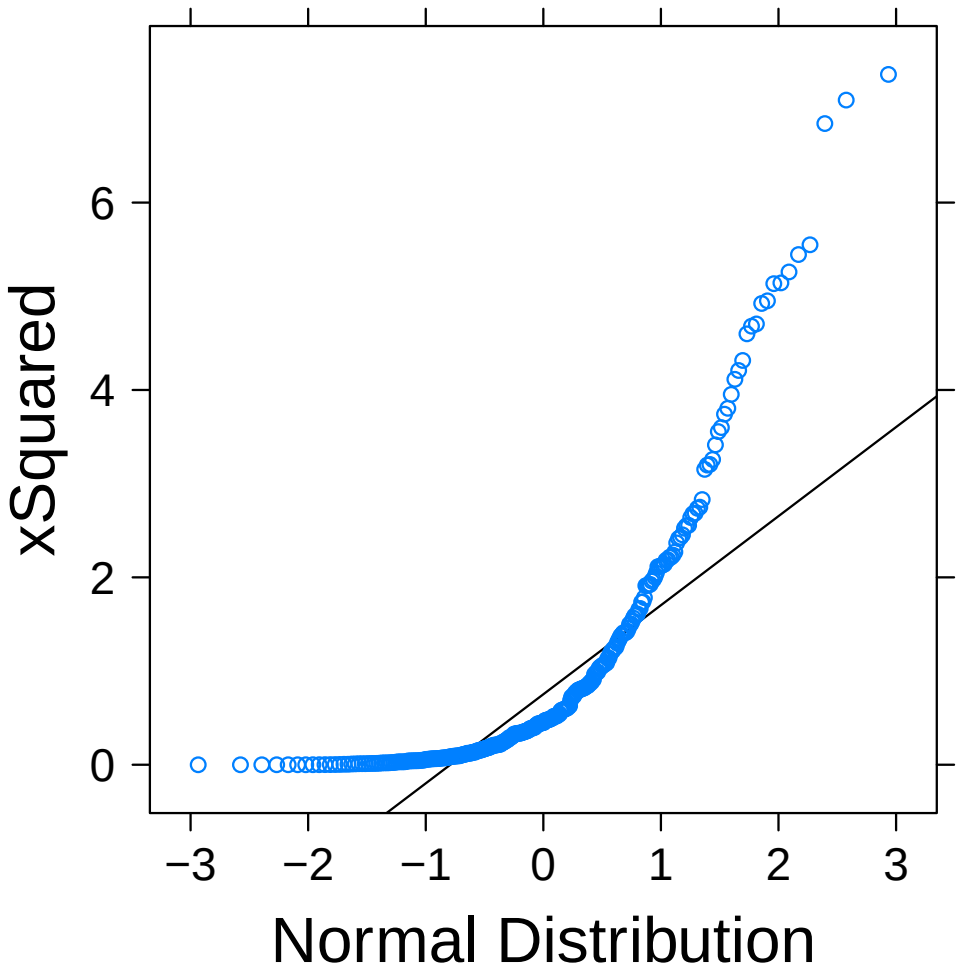
Power Transforms of Centered Data

- None of the examples from Cleveland use centered data.
- What happens when you power transform centered data?

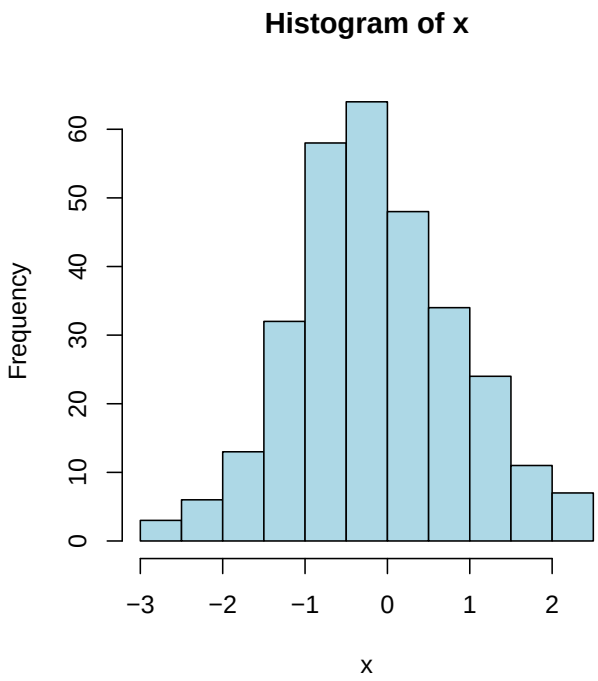
Power Transforms of Centered Data

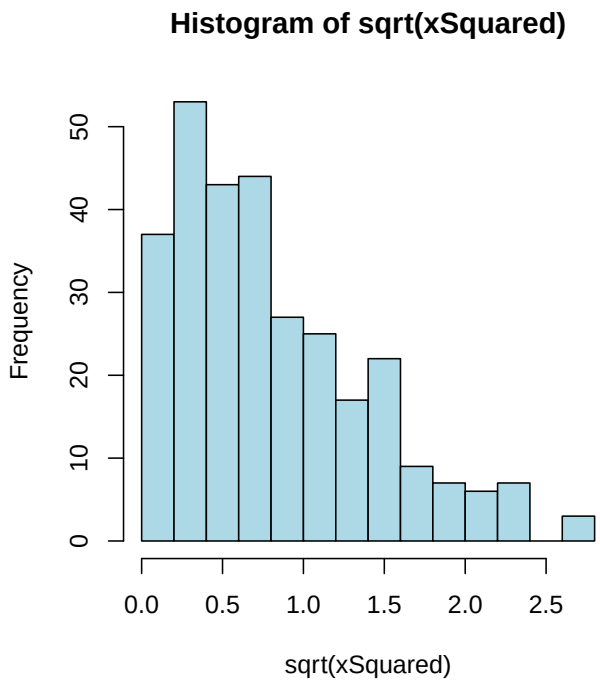


Power Transforms of Centered Data

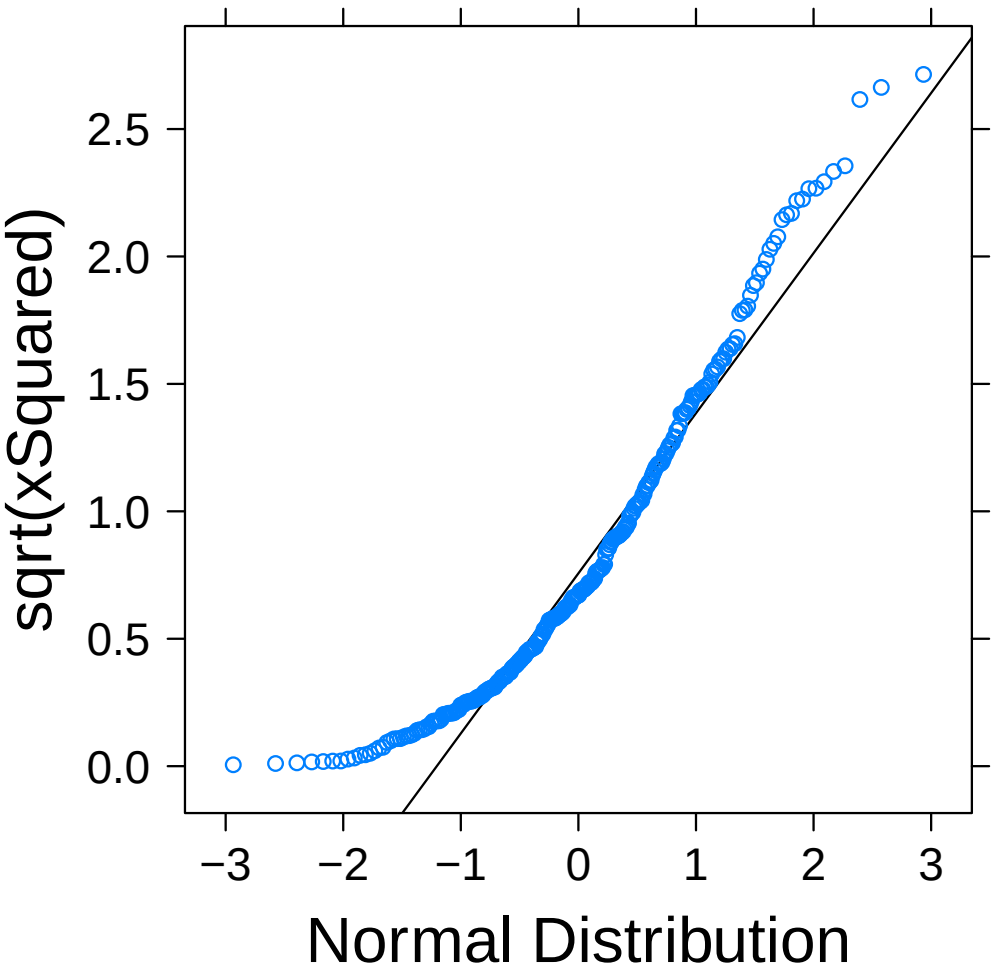


Power Transforms of Centered Data





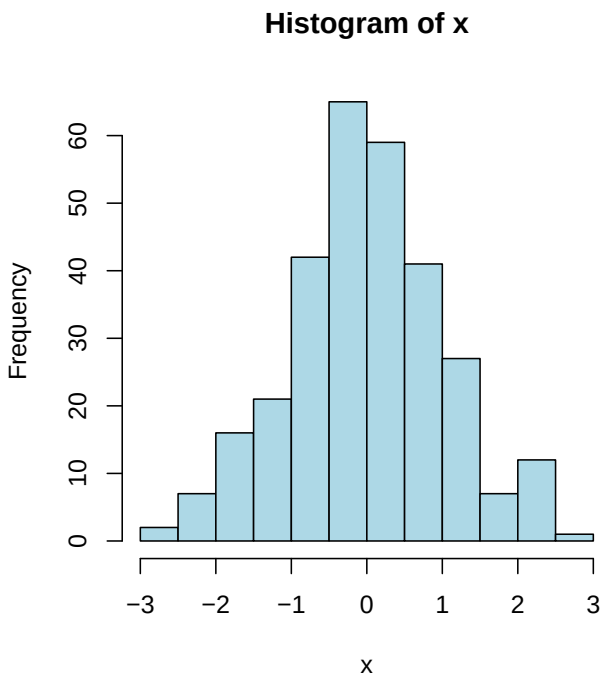
Power Transforms of Centered Data

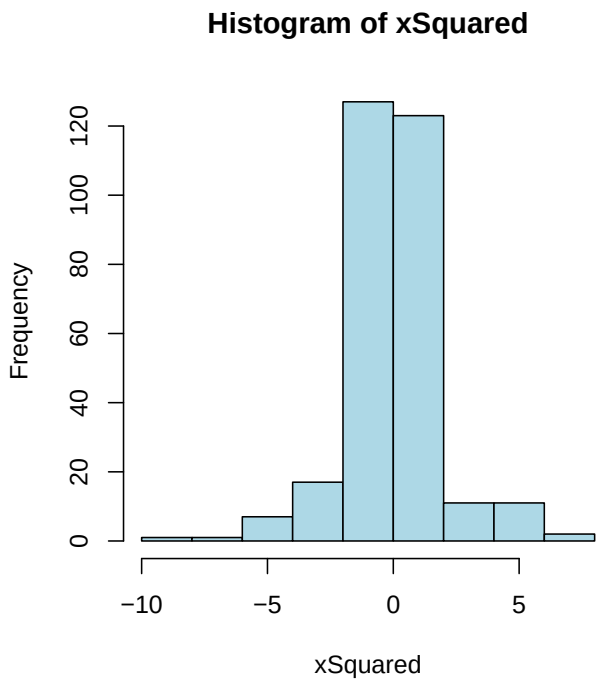


The centeredPower Function

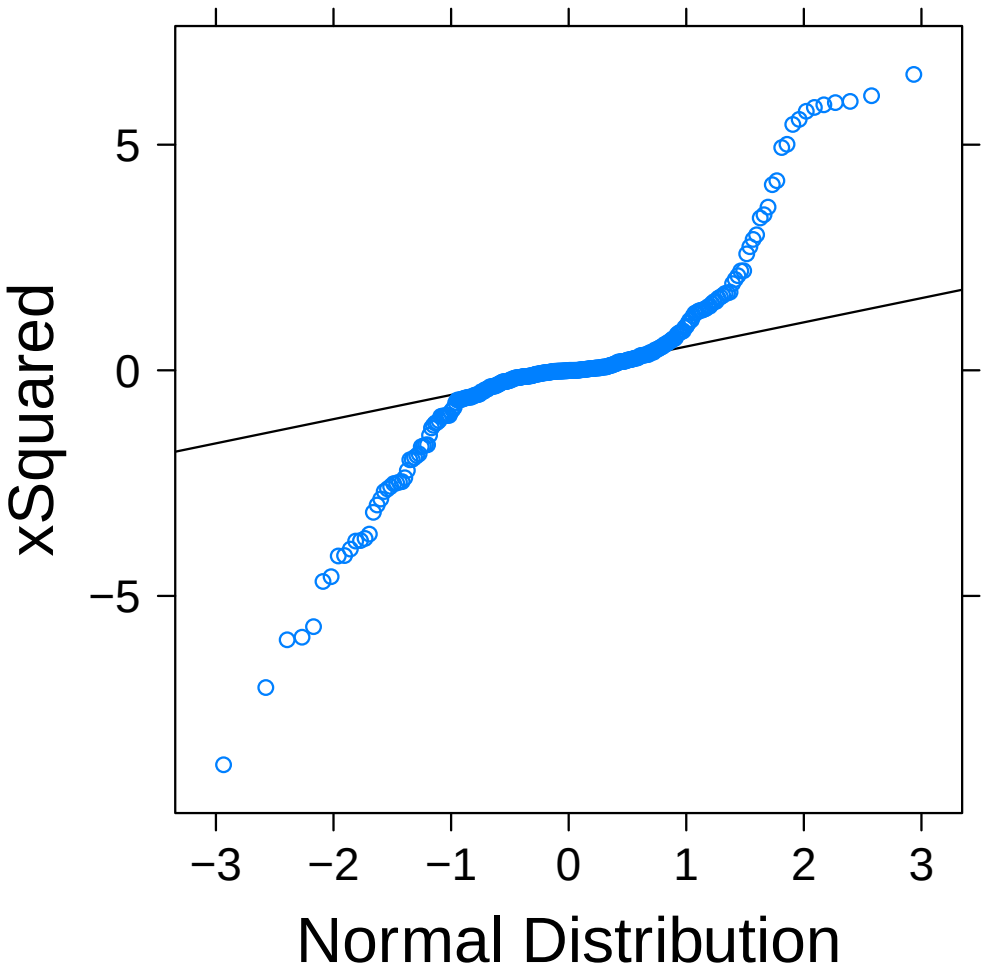
```
centeredPower <- function(dataVector, exponent) {
  tData <- dataVector
  tData[dataVector < 0] <- tData[dataVector < 0] * -1
  tData <- tData^(exponent)
  tData[dataVector < 0] <- tData[dataVector < 0] * -1
  return(tData)
}
```

centeredPower Transform of Centered Data

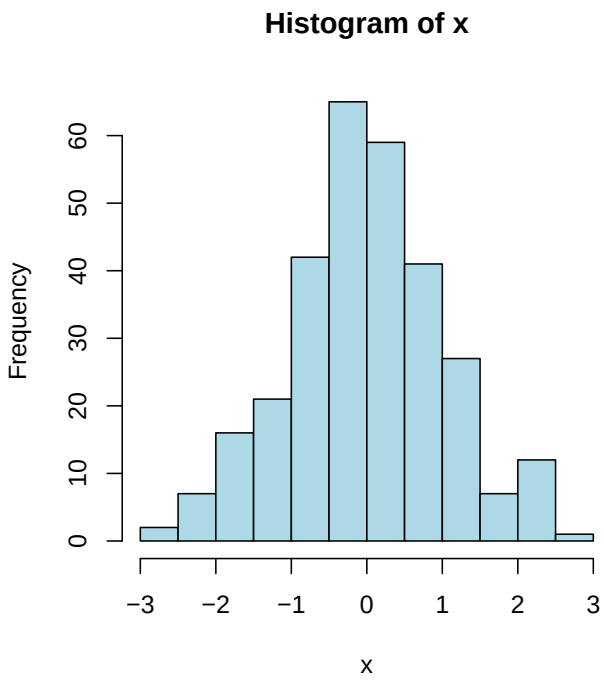


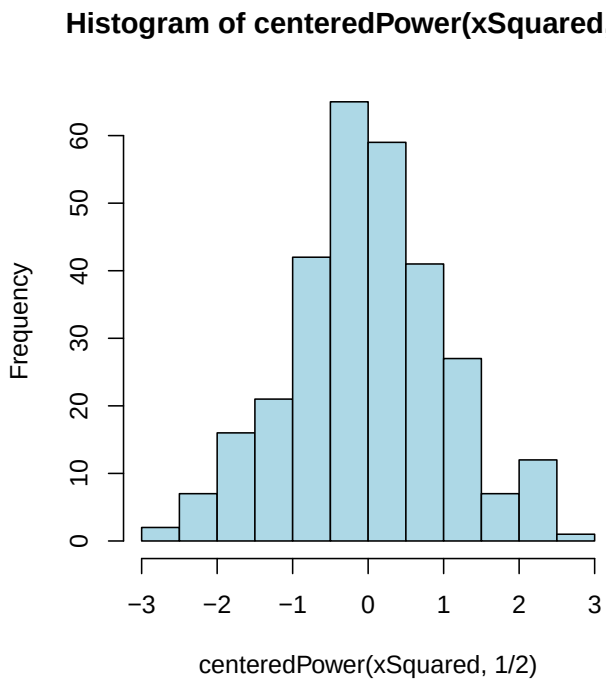


centeredPower Transform of Centered Data

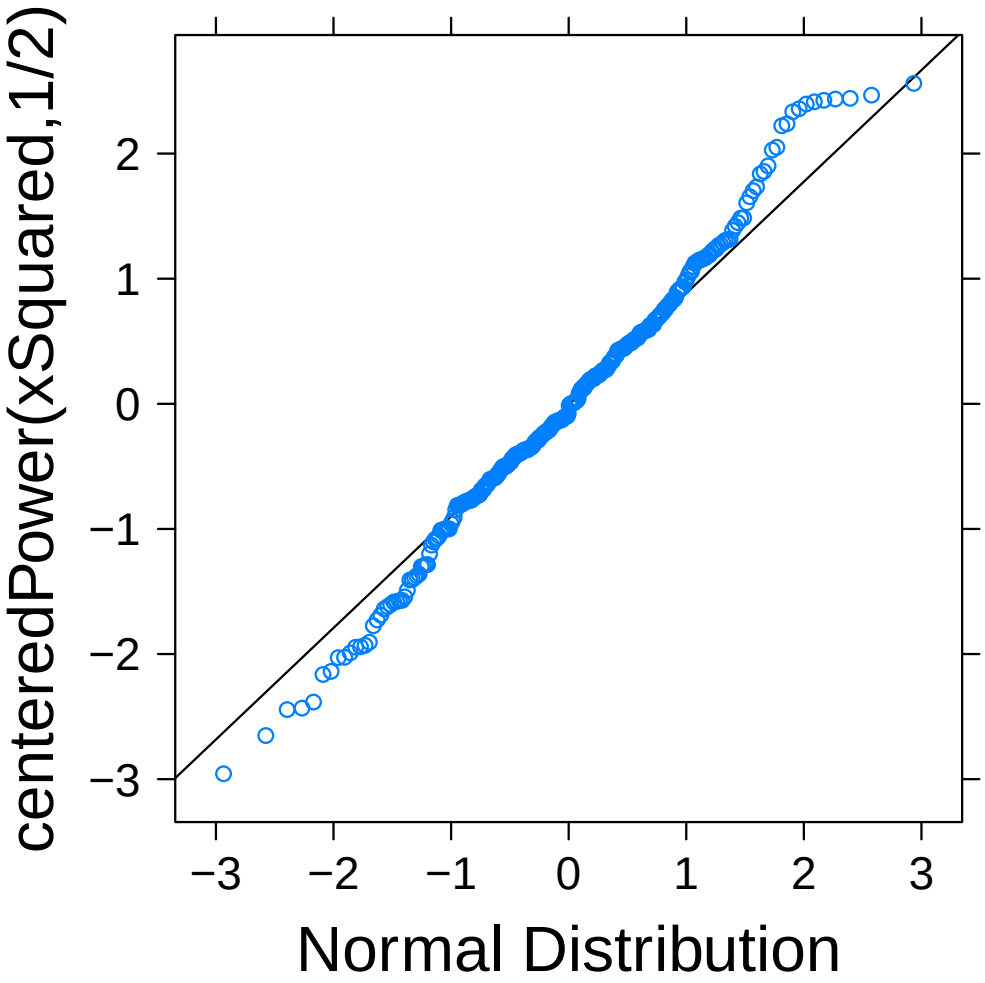


centeredPower Inverse Transform of Centered Data





centeredPower Inverse Transform of Centered Data



Next Week

- Autoregression and Cross–regression.
- Latent Growth Curves.
- Latent Differential Equations.