Lecture 8

Bill Perry

# Lecture 7: Review

Covered

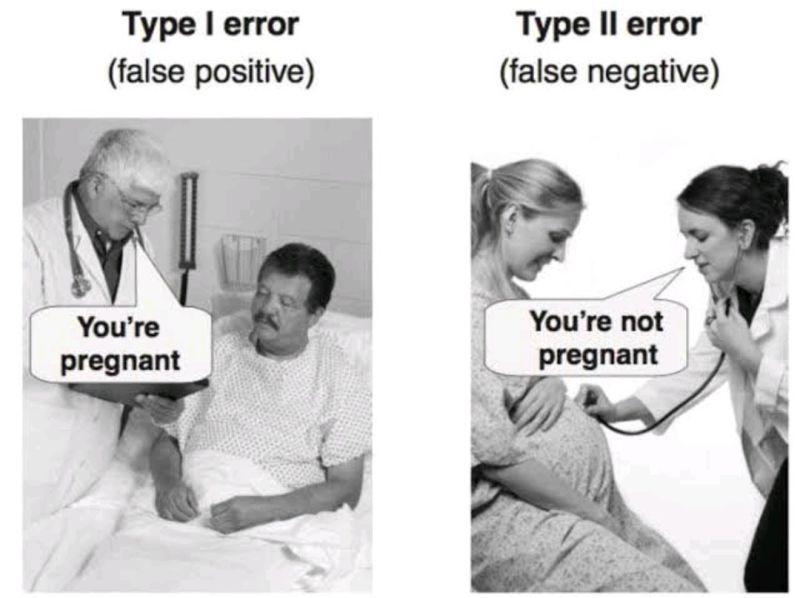
* Assumption tests for parametric tests
* Statistical vs Biological significance
* Nonparametric tests
  + Welch’s t-test: *when distribution normal but variance unequal*
  + Permutation test for two samples: *when distribution not normal (but both groups should still have similar distributions and ~equal variance)*
  + Mann-Whitney-Wilcoxon test: *when distribution not normal and/or outliers are present (but both groups should still have similar distributions and ~equal variance)*



# **Lecture 8:** Overview

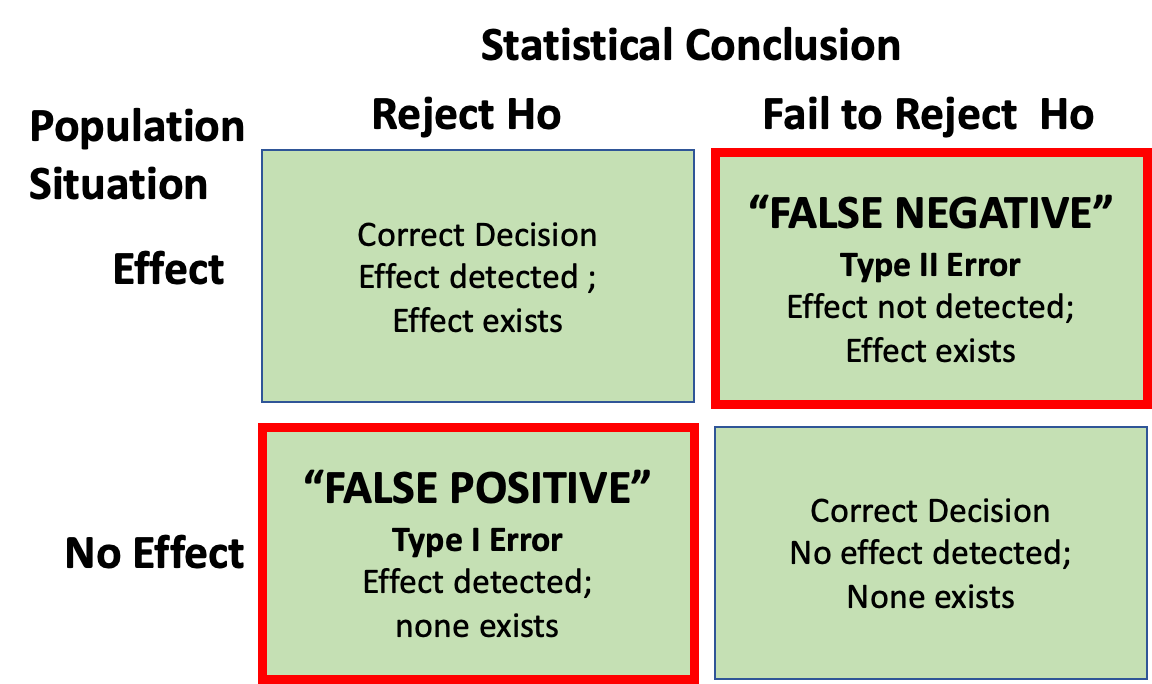
## The objectives:

* Decision errors
* Data exploration and transformation
* Exploratory graphical data analysis
* Graphical testing of assumptions
* Data transformation and standardization
* Outliers



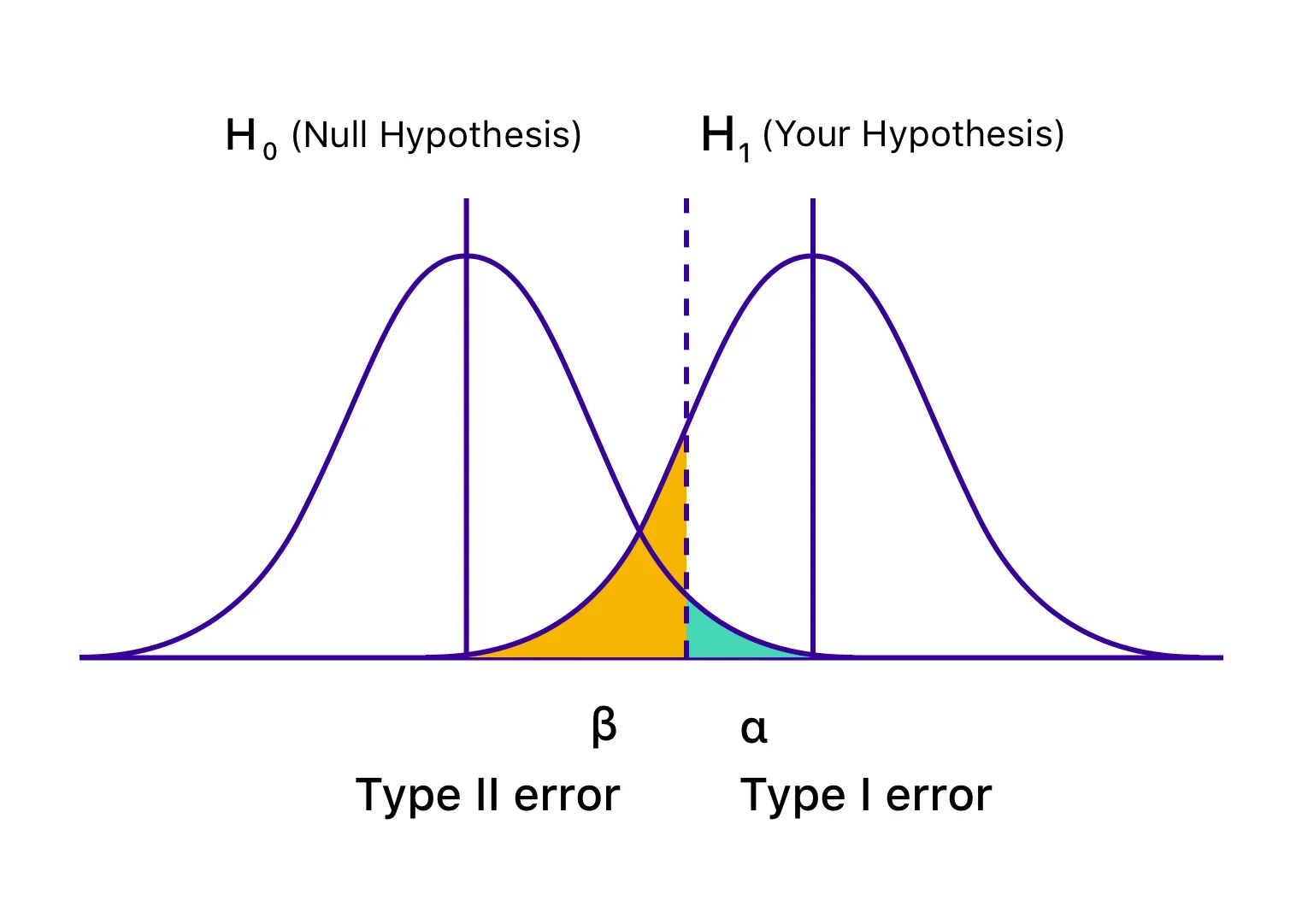
# **Decision errors**

* Even good studies can reach incorrect conclusions
* “Decision errors”
* Two types of decision errors
* Want to know probability of making these errors

<

# **Type I and Type II Errors**

* **Type I error rate**
  + **α**: wrongly reject H₀ when it’s true
  + α = 0.05 means a type I error rate of 5%
* **Type II error rate, β**
  + wrongly fail to reject H₀ when it’s false
* **Power = 1-β**: probability of correctly rejecting H₀ when H₁ is true
* Inverse relationship between type I and type II error - but not straightforward
* Result of chance - sample not representative of population
* Which type of error is more dangerous?

<

the dotted line is also the alpha = 0.05

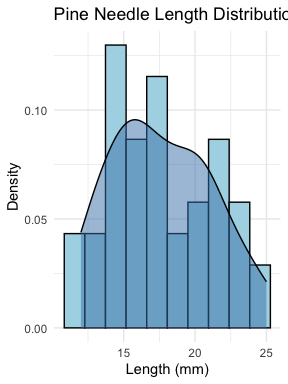
# **Exploratory graphical data analysis**

* Graphical exploration is one of first steps in data analysis:
  + Detect data entry errors
  + Pattern exploration
  + Assess assumptions of tests
  + Detect outliers
* Most important Q: shape of distribution?
* Determined by density plots: “density of different values”

# Let's examine our pine needle data  
pine\_data %>%   
 group\_by(wind) %>%  
 summarize(  
 n = n(),  
 mean = mean(len\_mm),  
 sd = sd(len\_mm),  
 min = min(len\_mm),  
 max = max(len\_mm)  
 )

# A tibble: 2 × 6  
 wind n mean sd min max  
 <chr> <int> <dbl> <dbl> <dbl> <dbl>  
1 lee 24 20.4 2.45 16 25  
2 wind 24 14.9 1.91 12 19

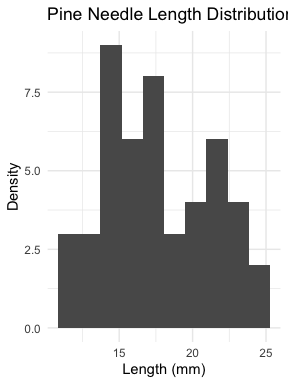
# Histogram with density  
ggplot(pine\_data, aes(x = len\_mm)) +  
 geom\_histogram(aes(y = ..density..),   
 fill = "lightblue",   
 color = "black",  
 bins = 10) +  
 geom\_density(alpha = 0.5, fill = "steelblue") +  
 labs(title = "Pine Needle Length Distribution",  
 x = "Length (mm)",   
 y = "Density") +  
 theme\_minimal()



# **Types of Exploratory Plots**

* **Histograms**: data broken into intervals, number of observations in each interval plotted on y-axis
  + Not great for small samples

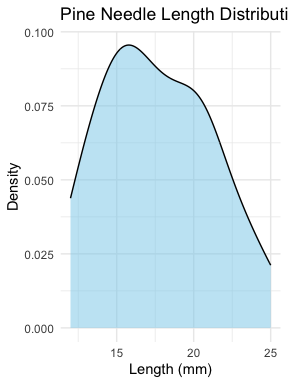
# Histogram with density  
ggplot(pine\_data, aes(x = len\_mm)) +  
 geom\_histogram(bins = 10) +  
 labs(title = "Pine Needle Length Distribution",  
 x = "Length (mm)",   
 y = "Density") +  
 theme\_minimal()



# **Types of Exploratory Plots**

* **Kernel density plot**: data broken into intervals, normal distribution assumed within each interval, sum of density functions plotted

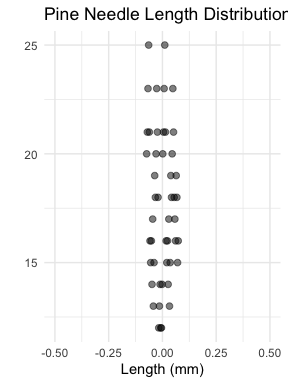
# Kernel density plot  
ggplot(pine\_data, aes(x = len\_mm)) +  
 geom\_density(fill = "skyblue", alpha = 0.5) +  
 labs(title = "Pine Needle Length Distribution",  
 x = "Length (mm)",   
 y = "Density") +  
 theme\_minimal()



# **Types of Exploratory Plots**

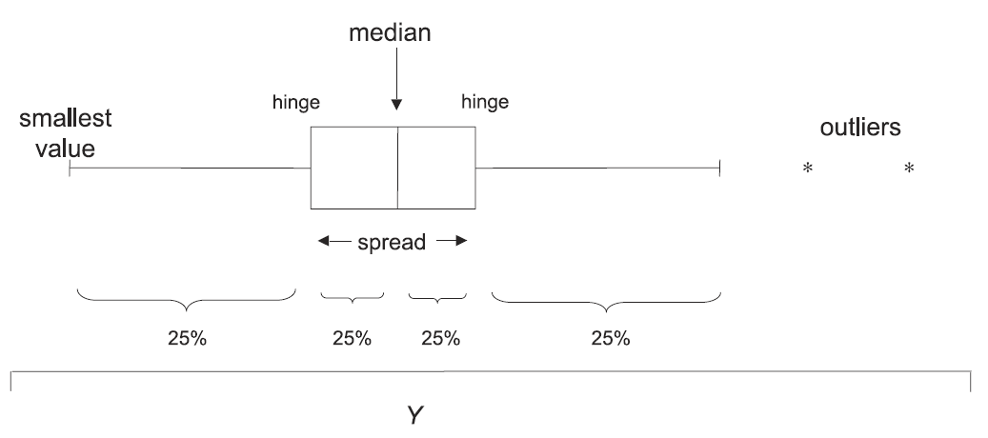
* **Dotplots**: each value represented as a dot along the measurement scale

# Dot plot of pine needle lengths  
ggplot(pine\_data, aes(x = 0, y = len\_mm)) +  
 geom\_point(size = 2, alpha = 0.5,  
 position = position\_dodge2(width=.15)) +  
 # geom\_jitter(width = 0.1, height = .05, size = 2, alpha = 0.5) +  
 labs(title = "Pine Needle Length Distribution",  
 x = "Length (mm)",   
 y = "") +  
 scale\_x\_continuous(limits = c(-.5, .5))+  
 theme\_minimal()

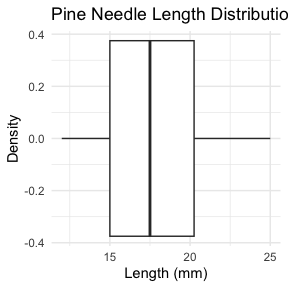


# **Types of Exploratory Plots**

* **Boxplot**: displays median, quartiles, range, outliers
  + Good when n > ~10

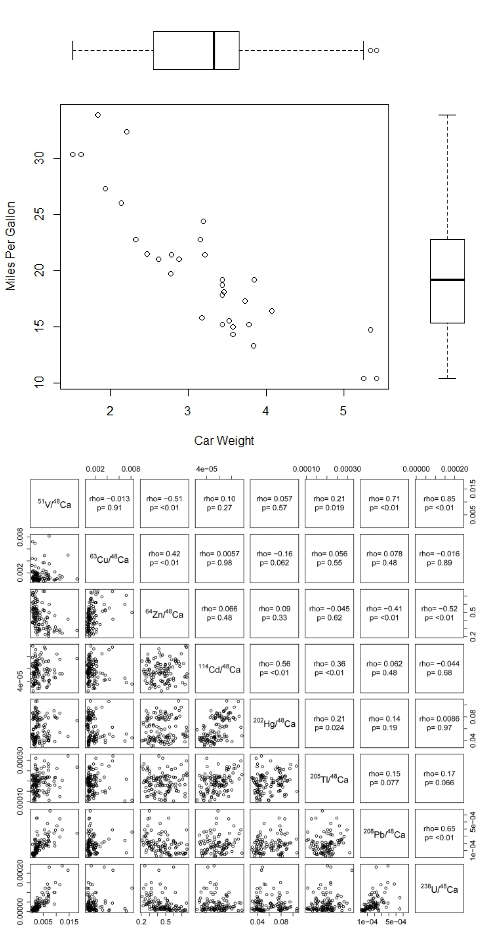


# Kernel density plot  
#| message: false  
#| warning: false  
#| fig-height: 4  
#| fig-width: 3  
#| include: true  
#| paged-print: false  
#|   
ggplot(pine\_data, aes(x = len\_mm)) +  
 geom\_boxplot()+  
 labs(title = "Pine Needle Length Distribution",  
 x = "Length (mm)",   
 y = "Density") +  
 theme\_minimal()



# **Types of Exploratory Plots**

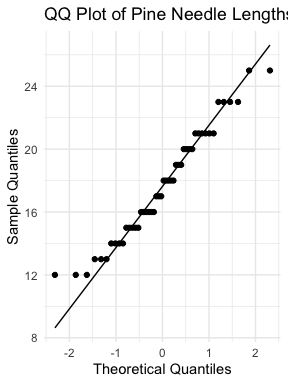
* **Scatter plot**: display of bivariate data
  + Shows distribution, outliers, non-linearity
* **Scatter matrix**: like scatterplot, but for multiple variables -will show later

<

# **Types of Exploratory Plots**

* **QQ plots**: compare quantiles of distribution against theoretical distribution (e.g. normal)

# qqplot  
# QQ plot for pine needle lengths  
ggplot(pine\_data, aes(sample = len\_mm)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 labs(title = "QQ Plot of Pine Needle Lengths",  
 x = "Theoretical Quantiles",   
 y = "Sample Quantiles") +  
 theme\_minimal()

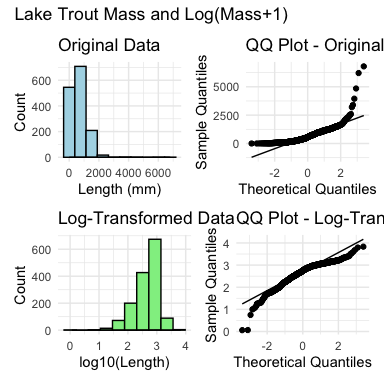


# ggplot(pine\_data, aes(sample = len\_mm)) +  
# stat\_qq(color = "darkgreen", size = 2, alpha = 0.6) +  
# stat\_qq\_line(color = "blue", linewidth = 1, linetype = "dashed") +  
# labs(title = "QQ Plot of Pine Needle Lengths",  
# x = "Theoretical Quantiles",   
# y = "Sample Quantiles") +  
# theme\_minimal()

# **Data transformation and standardization**

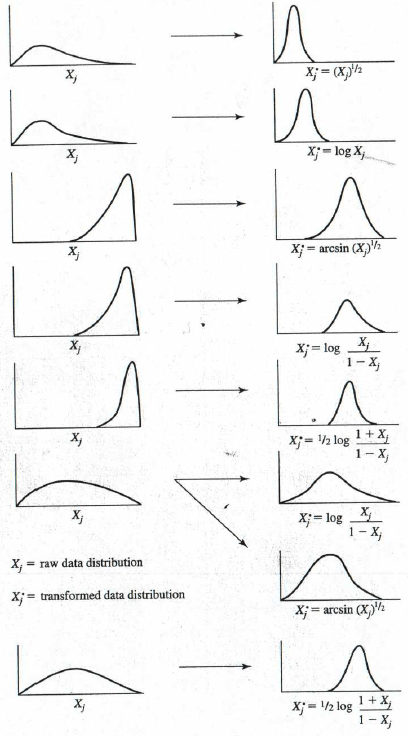
* If data don’t meet distributional assumptions can try transforming:
  + Approximate a normal distribution of data and errors
  + Improve homogeneity of variance
  + Reduce effect of outliers
  + Improve linearity for regression analysis
  + Reduce interactions between variables
* Data transformation changes the scale on which data are measured
* Common transformations:
  + Right-skewed data: power (root) transformations, log10 transformation
  + Left-skewed data: power transformations, log10 of (constant - x)
  + Percentages/proportions (bounded): Arcsine transformation
  + Rank transformation: most extreme, leads to loss of information

# Let's apply a log transformation to our pine needle data  
lt\_df <- read\_csv("data/lake\_trout.csv")  
  
# Let's apply a log transformation to our pine needle data  
lt\_df <- lt\_df %>%  
 mutate(log\_mass = log10(mass\_g +1))  
  
# Create before and after plots to show transformation effect  
lt\_hist\_1\_plot <- ggplot(lt\_df, aes(x = mass\_g)) +  
 geom\_histogram(bins = 10, fill = "lightblue", color = "black") +  
 geom\_density(alpha = 0.5) +  
 labs(title = "Original Data", x = "Length (mm)", y = "Count") +  
 theme\_minimal()  
  
lt\_qq\_1\_plot <- ggplot(lt\_df, aes(sample = mass\_g)) +  
 geom\_qq() +   
 geom\_qq\_line() +  
 labs(title = "QQ Plot - Original", x = "Theoretical Quantiles", y = "Sample Quantiles") +  
 theme\_minimal()  
  
lt\_hist\_2\_log\_plot <- ggplot(lt\_df, aes(x = log\_mass)) +  
 geom\_histogram(bins = 10, fill = "lightgreen", color = "black") +  
 geom\_density(alpha = 0.5) +  
 labs(title = "Log-Transformed Data", x = "log10(Length)", y = "Count") +  
 theme\_minimal()  
  
lt\_qq\_2\_log\_plot <- ggplot(lt\_df, aes(sample = log\_mass)) +  
 geom\_qq() +   
 geom\_qq\_line() +  
 labs(title = "QQ Plot - Log-Transformed", x = "Theoretical Quantiles", y = "Sample Quantiles") +  
 theme\_minimal()  
  
# Combine plots  
(lt\_hist\_1\_plot + lt\_qq\_1\_plot) / (lt\_hist\_2\_log\_plot + lt\_qq\_2\_log\_plot)+  
 plot\_annotation(  
 title = "Lake Trout Mass and Log(Mass+1)"  
 )



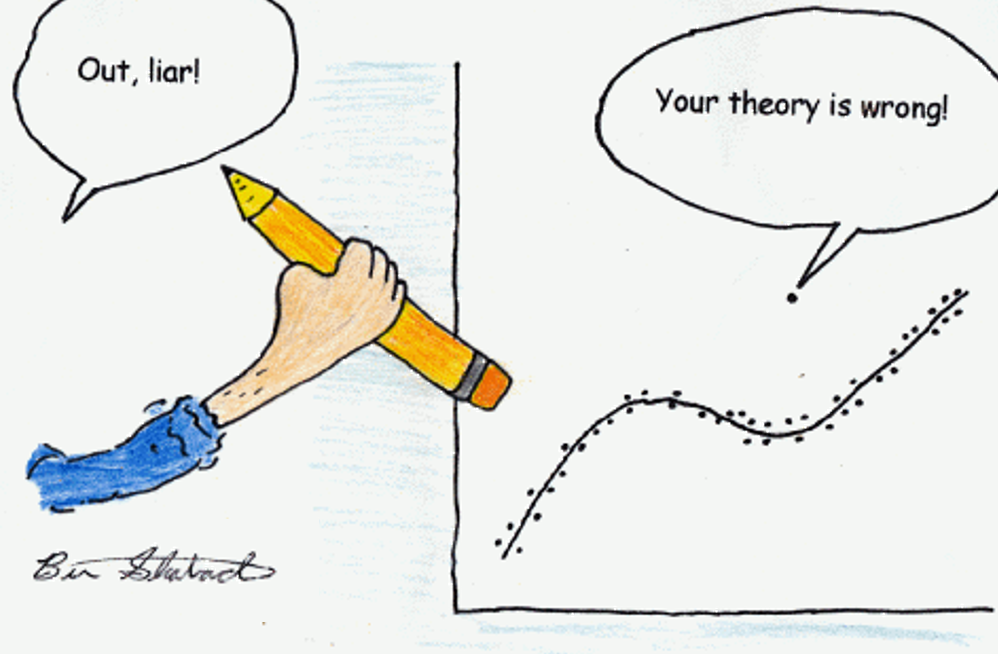
# **Data transformation and standardization**

* Common transformations:
  + Right-skewed data: power (root) transformations, log10 transformation
  + Left-skewed data: power transformations, log10 of (constant - x)
  + Percentages/proportions (bounded): Arcsine transformation
  + Rank transformation: most extreme, leads to loss of information

<

# **Outliers**

* Outliers: unusual values that are outside the range of most other observations
  + Can significantly affect results of analysis
* Outliers identified using:
  + Formal tests (Dixon’s Q, Cook’s D)
  + Graphically, using boxplots or QQ plots
* What to do with outliers? Depends why they happened:
  + If obvious data entry error, can be removed
  + If part of the data:
    - Rerun analysis with and without outliers, report both results
    - Use tests robust to outliers or transform data
  + Unethical to remove inconvenient outliers

<

# **Final Activity:** Take home messages

**Common assumptions for tests:**

1. Normality: Data comes from normally distributed populations
2. Equal variances (for two-sample tests)
3. Independence: Observations are independent
4. No outliers: Extreme values can influence results

What can we do if our data violates these assumptions?

Alternatives

* Data transformation (log, square root, etc.)
* Non-parametric tests
* Bootstrapping approaches

# **Summary and Conclusions**

In this activity, we’ve:

1. Explored decision errors (Type I and Type II) and their implications
2. Learned various methods for exploratory data analysis
3. Discussed data transformations to meet statistical assumptions
4. Examined approaches for handling outliers

**Key takeaways:**

* Always explore your data visually before formal analysis
* Consider the assumptions of statistical tests and check if they are met
* Choose appropriate transformations or alternative tests when assumptions are violated
* Be transparent about handling outliers and report all analytical decisions

# **What do you see as the key points?**

Things that stood out

# **What are the muddy points?**

What does not make sense or what questions do you have…

What makes you nervous?