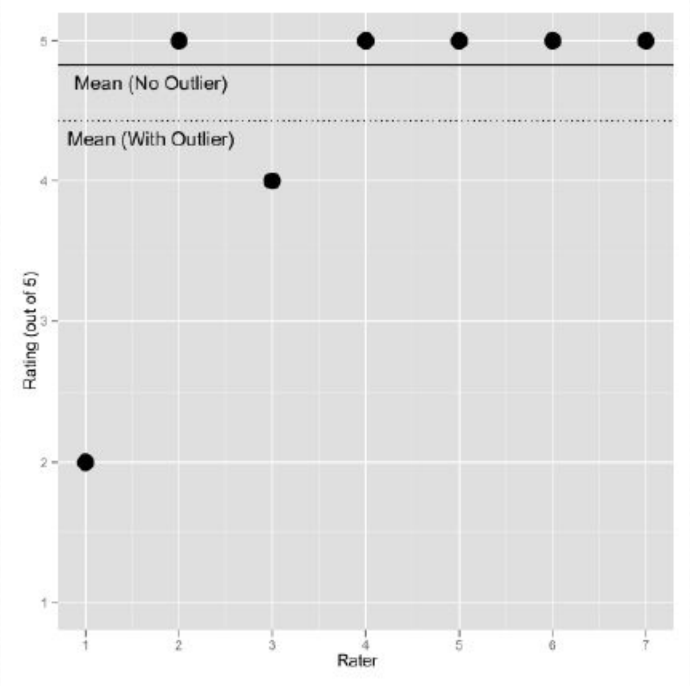
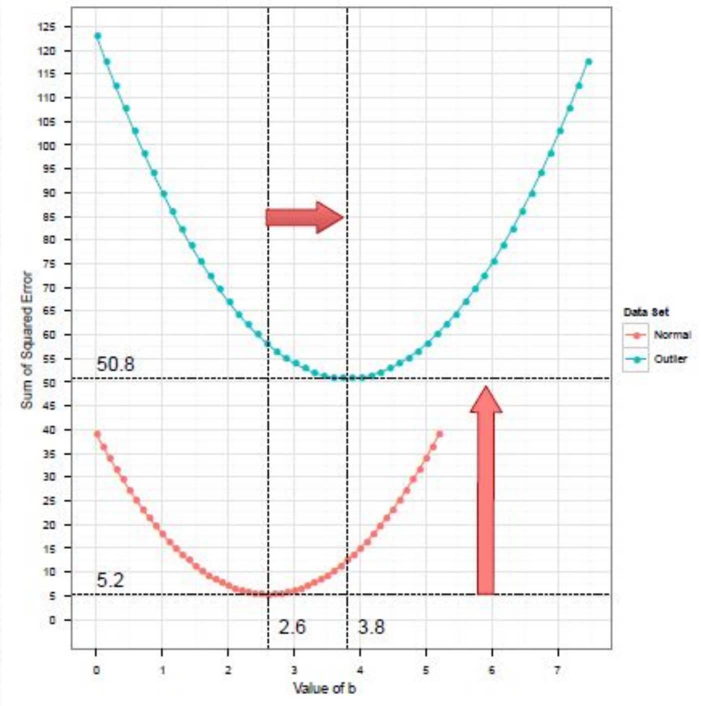
Session 5:

Exploring Assumptions

Normality and Homogeneity of Variance

Outliers Impact





Assumptions

Parametric tests based on the normal distribution assume:

- Additivity and linearity
- Normality something or other
- Homogeneity of Variance
- Independence

Additivity and Linearity

- The outcome variable is, in reality, linearly related to any predictors.
- If you have several predictors then their combined effect is best described by adding their effects together.
- If this assumption is not met then your model is invalid.

Normality Something or Other

The normal distribution is relevant to:

- Parameters
- Confidence intervals around a parameter
- Null hypothesis significance testing

This assumption tends to get incorrectly translated as 'your data need to be normally distributed'.

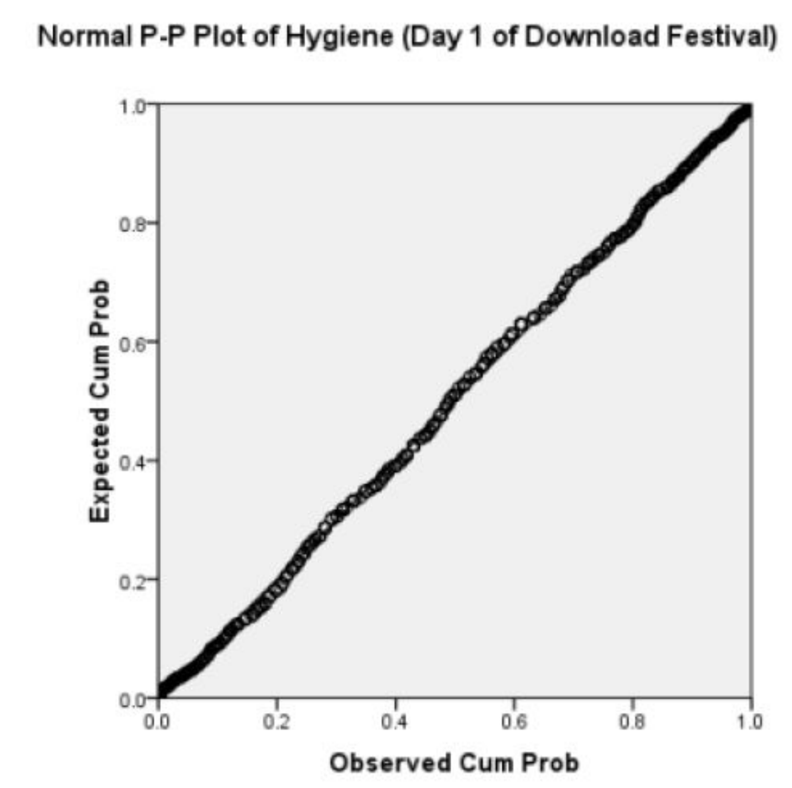
When does the Assumption of Normality Matter?

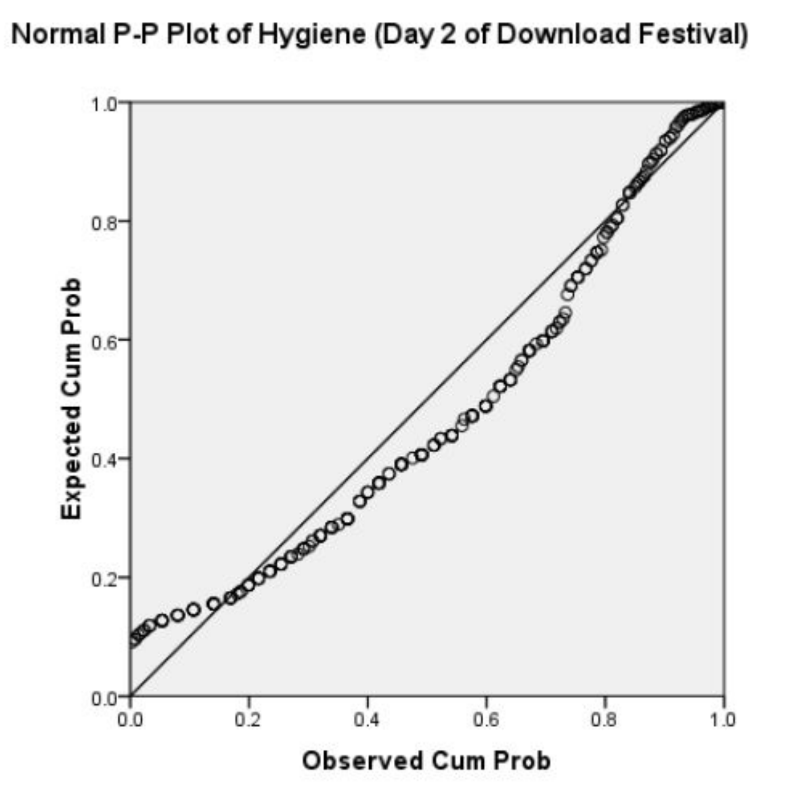
- In small samples – The central limit theorem allows us to forget about this assumption in larger samples.
- In practical terms, as long as your sample is fairly large, outliers are a much more pressing concern than normality.

Spotting Normality

- We don't have access to the sampling distribution so we usually test the observed data
- Central Limit Theorem
  - If *N* > 30, the sampling distribution is normal anyway
- Graphical displays
  - P-P Plot (or Q-Q plot)
  - Histogram
- Values of Skew/Kurtosis
  - 0 in a normal distribution
  - Convert to z (by dividing value by *SE*)
- Kolmogorov-Smirnov Test
  - Tests if data differ from a normal distribution
  - Significant = non-Normal data
  - Non-Significant = Normal data

The P-P Plot


Normal P-P Plot of Hygiene (Day 1 of Download Festival)
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0


Normal P-P Plot of Hygiene (Day 2 of Download Festival)
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

Assessing Skew and Kurtosis

Skewness	-.004	1.095	1.033
Std. Error of Skewness	.086	.150	.218
Kurtosis	-.410	.822	.732
Std. Error of Kurtosis	.172	.299	.433
Range	3.67	3.44	3.39

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Hygiene (Day 1 of Download Festival)	.083	810	.000	.654	810	.000

a. Lilliefors Significance Correction

H_0 : Normality can be assumed.

H_1 : Normality cannot be assumed.

K-S Test

D(810)=0.083, sig=0.000(<0.05)

This test is significant

Reject H_0

Conclusion : It is not a normal distribution.

Homoscedasticity/ Homogeneity of Variance

- When testing several groups of participants, samples should come from populations with the same variance.
- In correlational designs, the variance of the outcome variable should be stable at all levels of the predictor variable.
- Can affect the two main things that we might do when we fit models to data:
  - Parameters
  - Null Hypothesis significance testing

Assessing Homoscedasticity/ Homogeneity of Variance

Graphs (see lectures on regression)

Levene's Tests

- Tests if variances in different groups are the same.
- Significant = Variances not equal
- Non-Significant = Variances are equal

Variance Ratio

- With 2 or more groups
- VR = Largest variance/Smallest variance
- If VR < 2, homogeneity can be assumed.

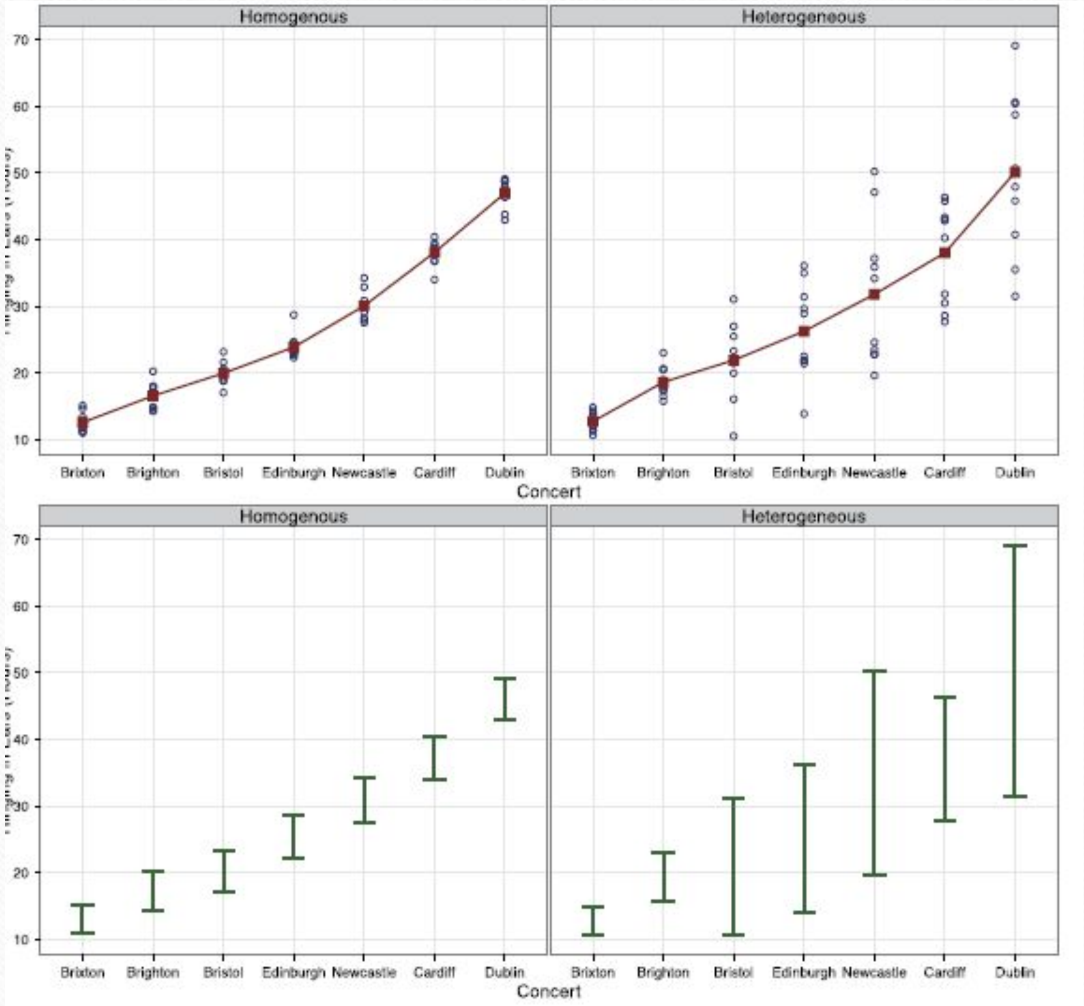
Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Age of surveyer	Based on Mean	.985	2	56	.380
	Based on Median	.499	2	56	.610
	Based on Median and with adjusted df	.499	2	36.562	.611
	Based on trimmed mean	.872	2	56	.424

H_0: *Homogeneity of Variance can be assumed.*
H_1: *Homogeneity of Variance cannot be assumed.*

Levene's Test
F(2.56)=0.985, sig =0.380 (>0.05)
This test is non-significant.
Accept H_0
Conclusion : The variances are about the same in different groups

Homogeneity of Variance



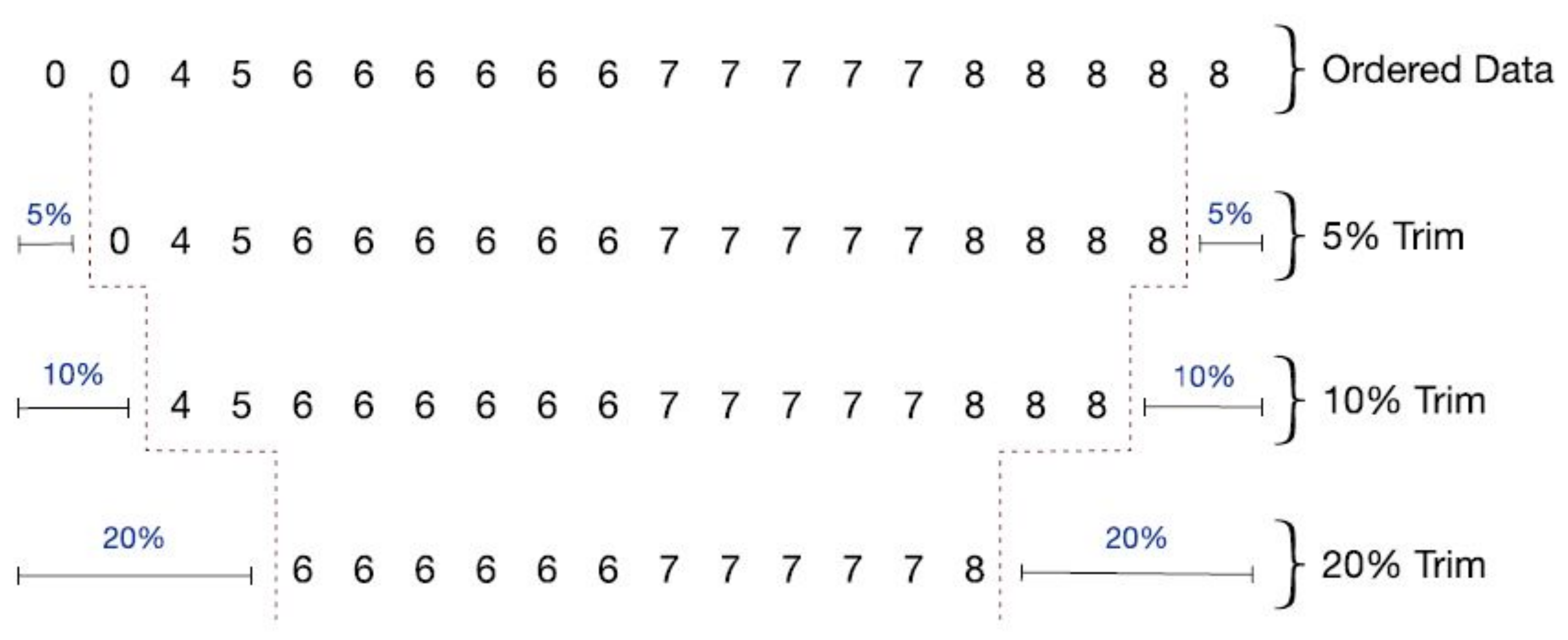
Independence

- The errors in your model should not be related to each other.
- If this assumption is violated: Confidence intervals and significance tests will be invalid.

Reducing Bias

- **Trim the data**: Delete a certain amount of scores from the extremes.
- **Windsorizing**: Substitute outliers with the highest value that isn't an outlier
- **Analyze with Robust Methods**: Bootstrapping
- **Transform the data:** By applying a mathematical function to scores

Trimming the Data



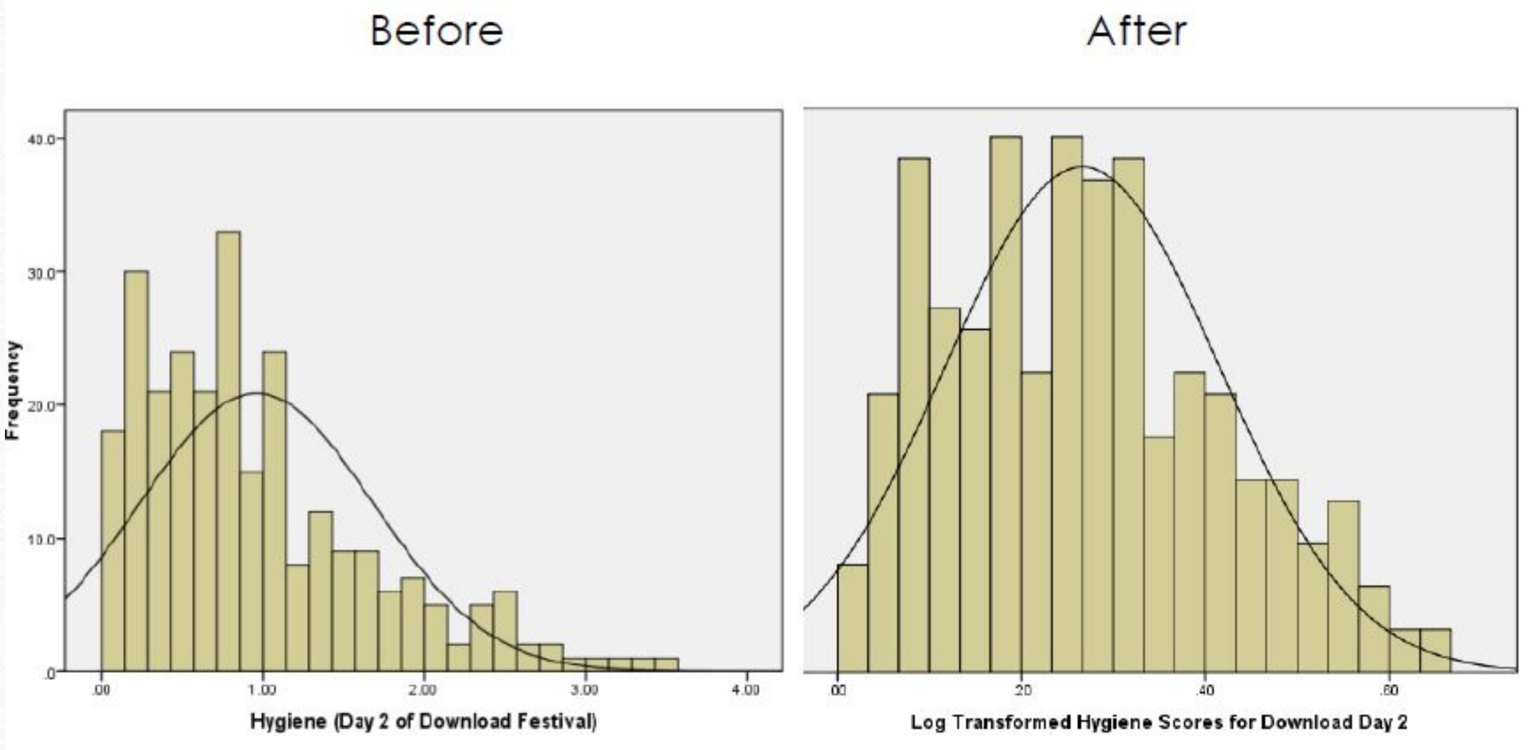
Robust Methods

	Comparing Treatments	Relationships
Principle	Bootstrap	Bootstrap
	Trimmed Means	Least Trimmed Squares
	M-estimators	M-estimators
	Median	Least Median of Squares
Equivalent Tests	T-test	Correlation
	ANOVA (Including factorial)	Regression
	ANCOVA	ANCOVA
	MANOVA	

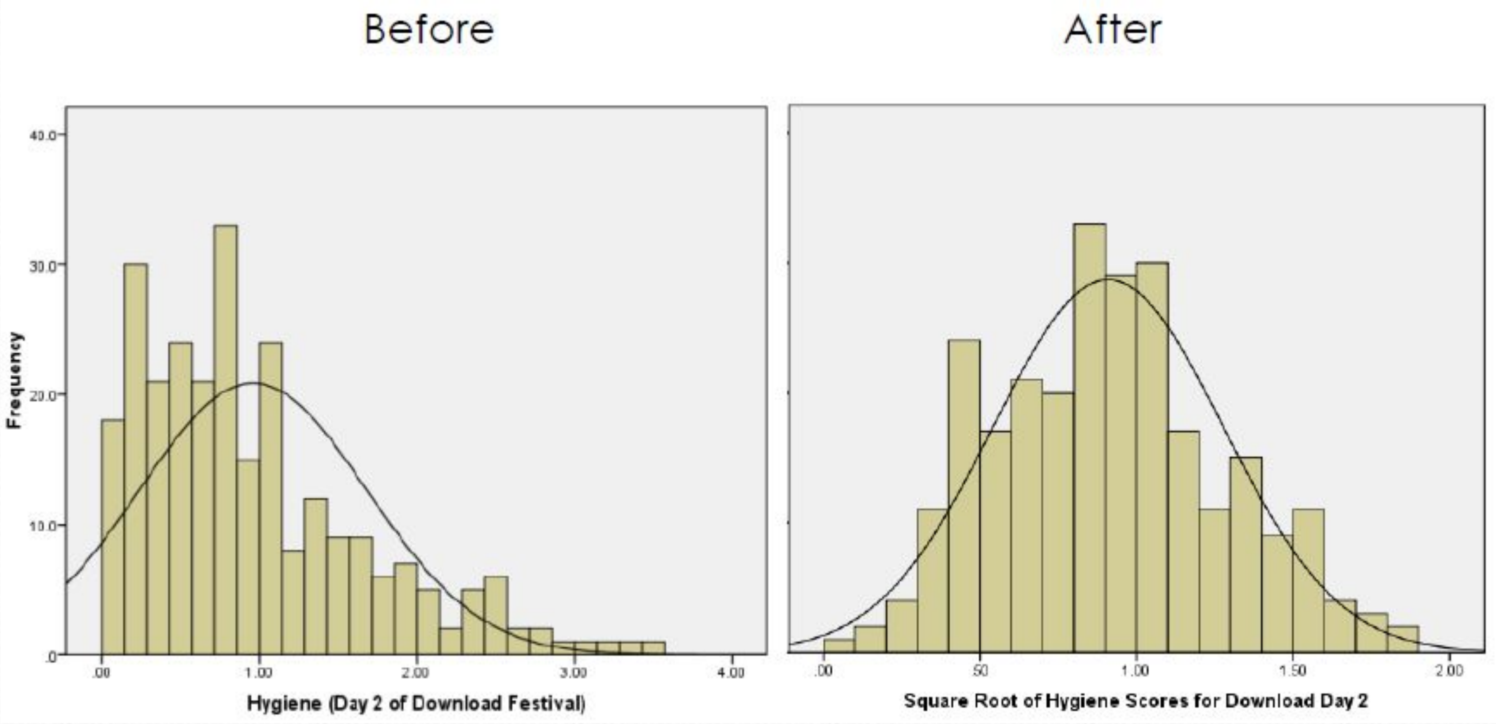
Transforming Data

- **Log Transformation** ($\log(x_i)$): Reduce positive skew.
- **Square Root Transformation** ($\sqrt{x_i}$): Also reduces positive skew. Can also be useful for stabilizing variance.
- **Reciprocal Transformation** ($1/x_i$): Dividing 1 by each score also reduces the impact of large scores. This transformation reverses the scores, you can avoid this by reversing the scores before the transformation, $1/(x_{Highest}-x_i)$.

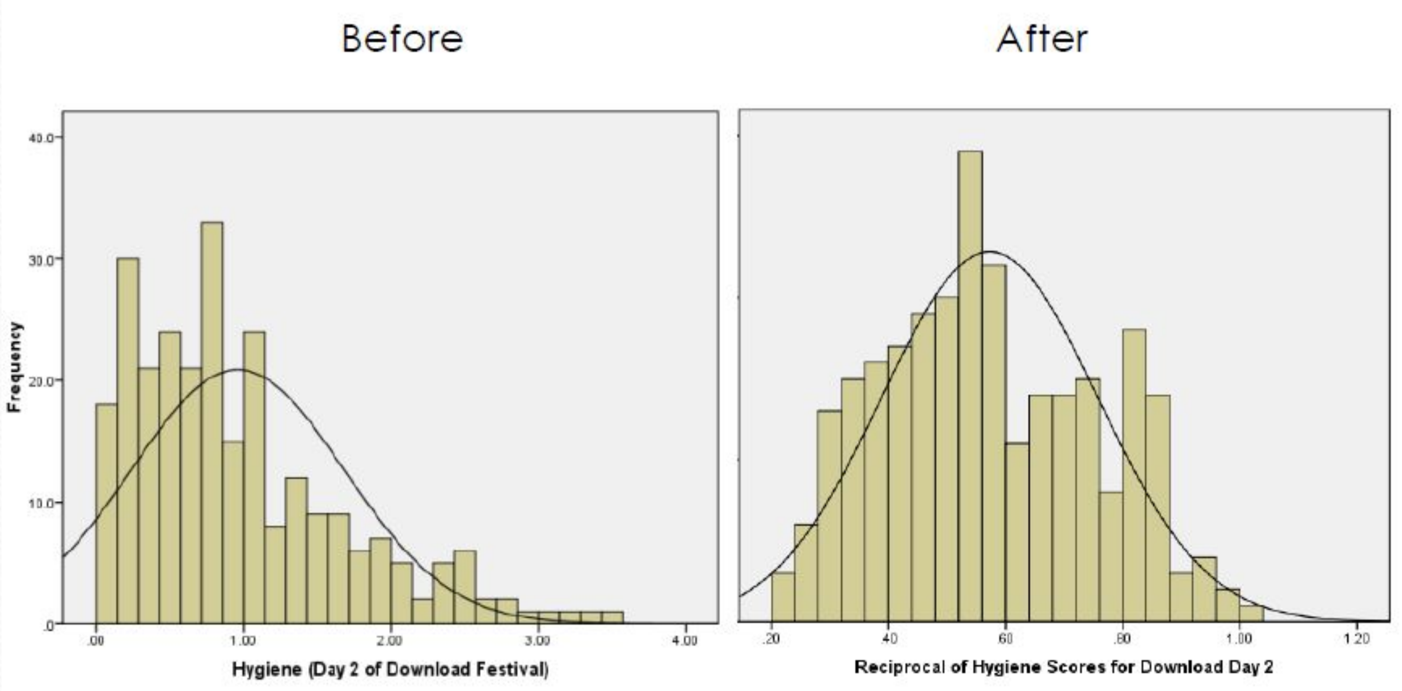
Log Transformation



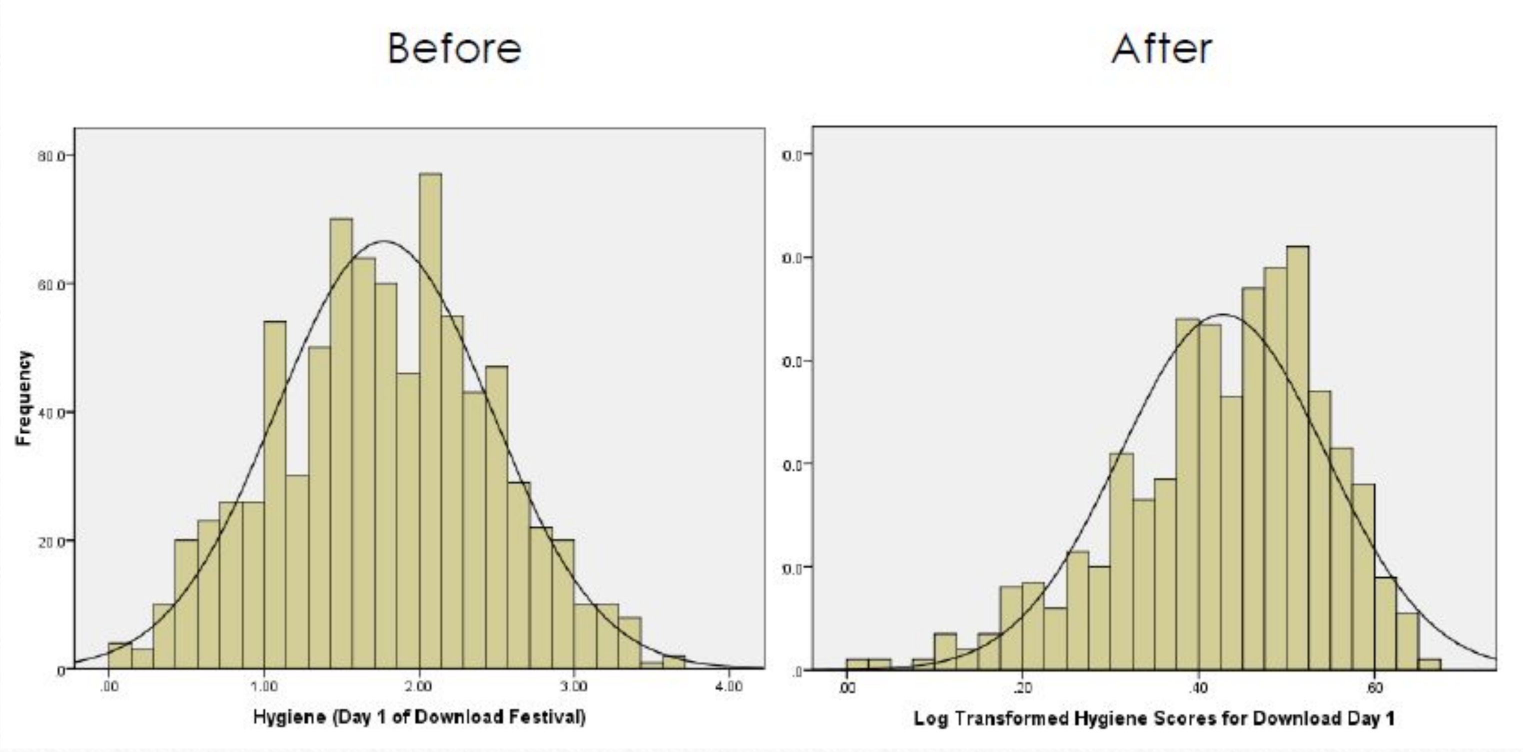
Square Root Transformation



Reciprocal Transformation



But …


Before
After
Frequency
80.0
60.0
40.0
20.0
0
.00
1.00
2.00
3.00
4.00
Hygiene (Day 1 of Download Festival)
.00
.20
.40
.60
Log Transformed Hygiene Scores for Download Day 1

To Transform … Or Not

Transforming the data helps as often as it hinders the accuracy of *F* (Games & Lucas, 1966).

Games (1984):

- The central limit theorem: sampling distribution will be normal in samples > 40 anyway.
- Transforming the data changes the hypothesis being tested
- E.g. when using a log transformation and comparing means you change from comparing arithmetic means to comparing geometric means
- In small samples it is tricky to determine normality one way or another.
- The consequences for the statistical model of applying the 'wrong' transformation could be worse than the consequences of analysing the untransformed scores.