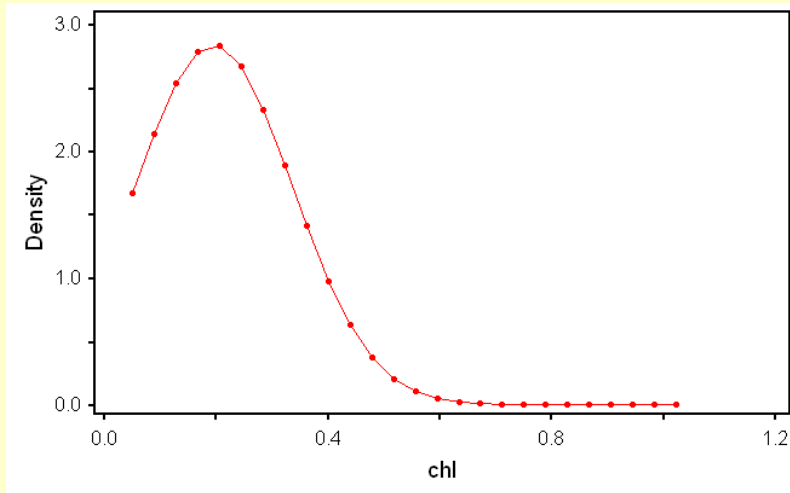
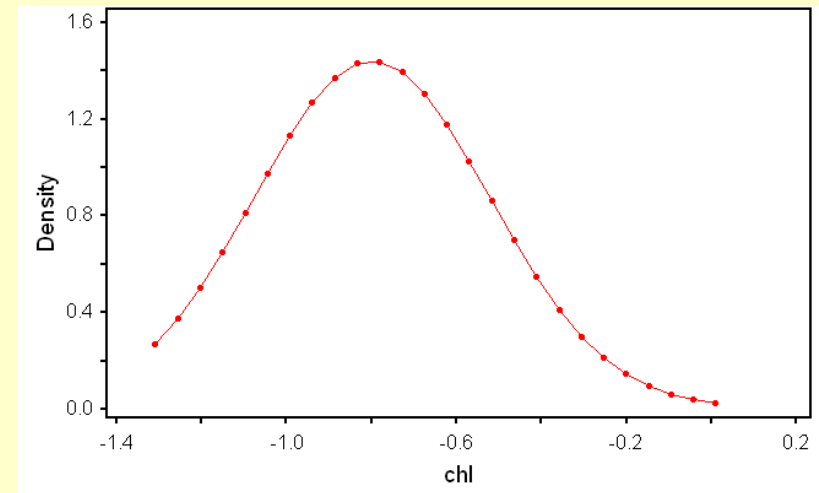


Data Distributions and Transformations



Chl Concentration



Log (Chl Conc.)

Approach

- We can transform the data to achieve normality
- Need to implement monotonic transformations:

Actual Values Change

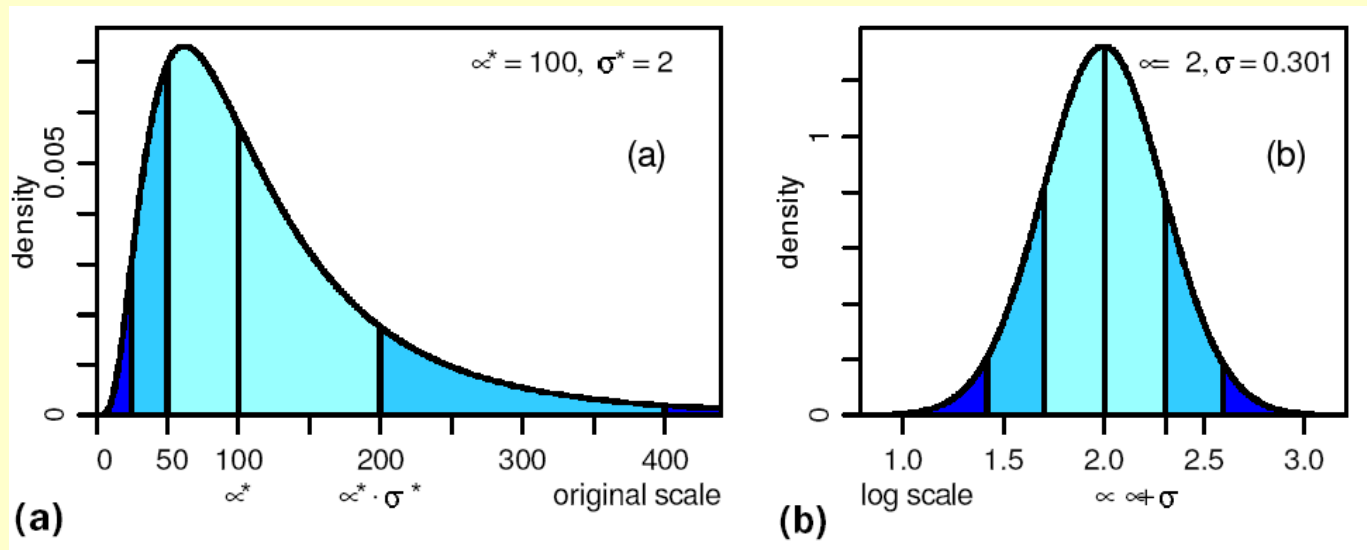
Ranks Do Not Change

The Lognormal Distribution

$$\text{Log } X \sim N(\mu, \sigma)$$

A probability distribution of a random variable whose logarithm is normally distributed.

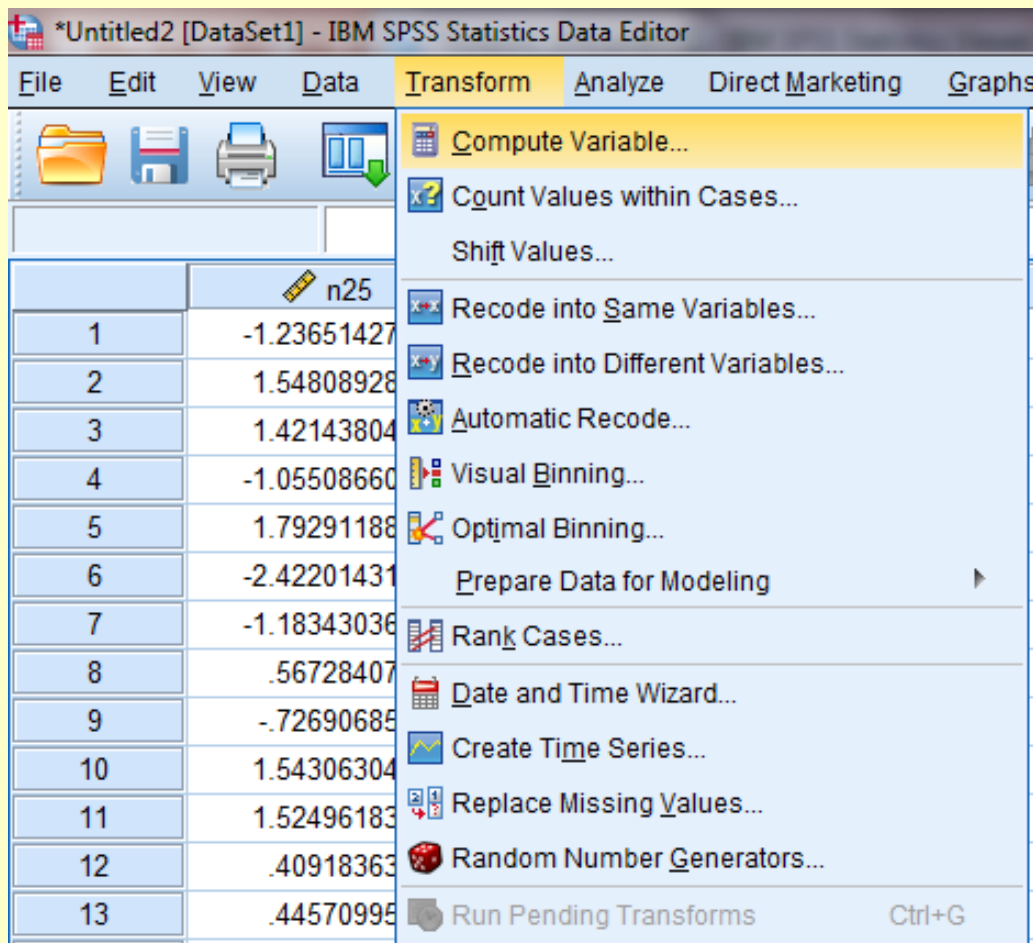
Like Normal Distribution, described using two parameters:
Mean and S.D.



(Limpert et al. 2001)

Transforming Data in SPSS

Transform / Compute Variable...

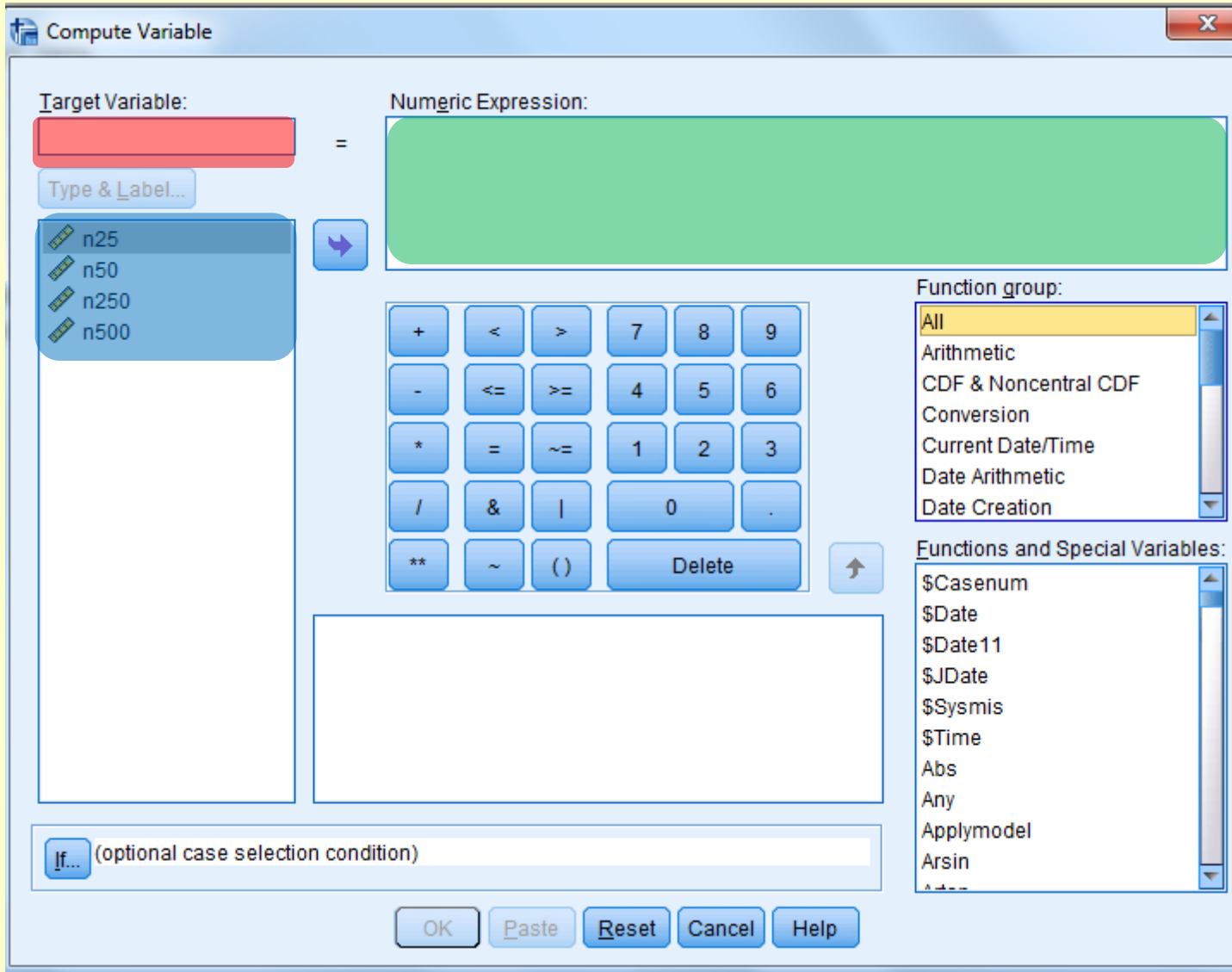


We compute and name a new variable

New variable is added to the existing dataset

Save / export dataset with changes

Transforming Data in SPSS



Target:
The new variable you are creating

Variable:
The old variable

Expression:
What you are doing to the variable

Monotonic Data Transformations

Reasonable and acceptable
domain of x

Range of $f(x)$

MONOTONIC TRANSFORMATIONS

(x)



$f(x)$

x^0 (power)

all

0 or 1 only

P / A

$x^{1/2}$ (power)

nonnegative

nonnegative

Power exponents:

$\frac{1}{2}$ power (square root)

Square root transform is similar, but less drastic than the log transform.

Special treatment of zeros not necessary

Note: 0 power transformation is NOT really monotonic

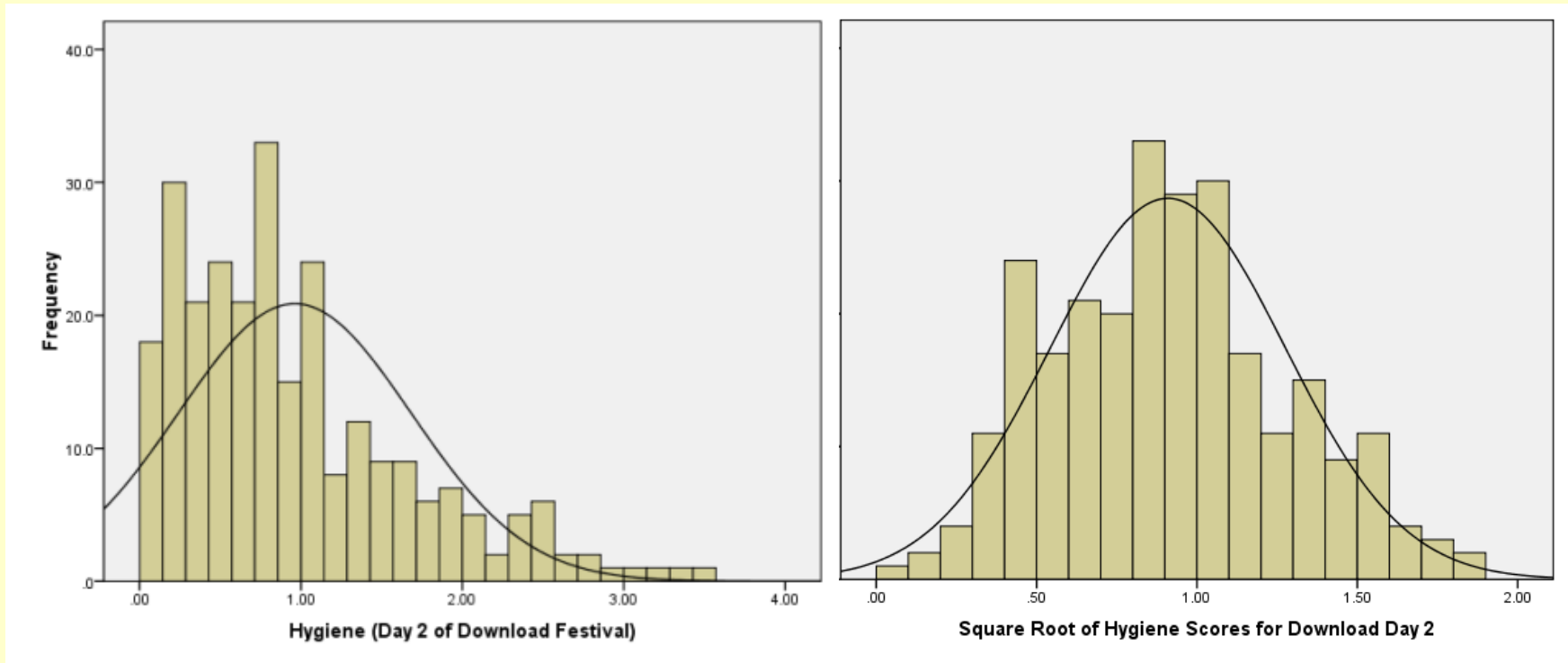
It recodes data as Presence / Absence (0 / 1)

Square Root Transformation

Square Root Transformation ($\sqrt{X_i}$):
Reduces positive skew. Useful for stabilizing variance

Before

After



Monotonic Data Transformations

Logarithmic transformation $f(x) = \ln(x)$ OR $\log(x)$

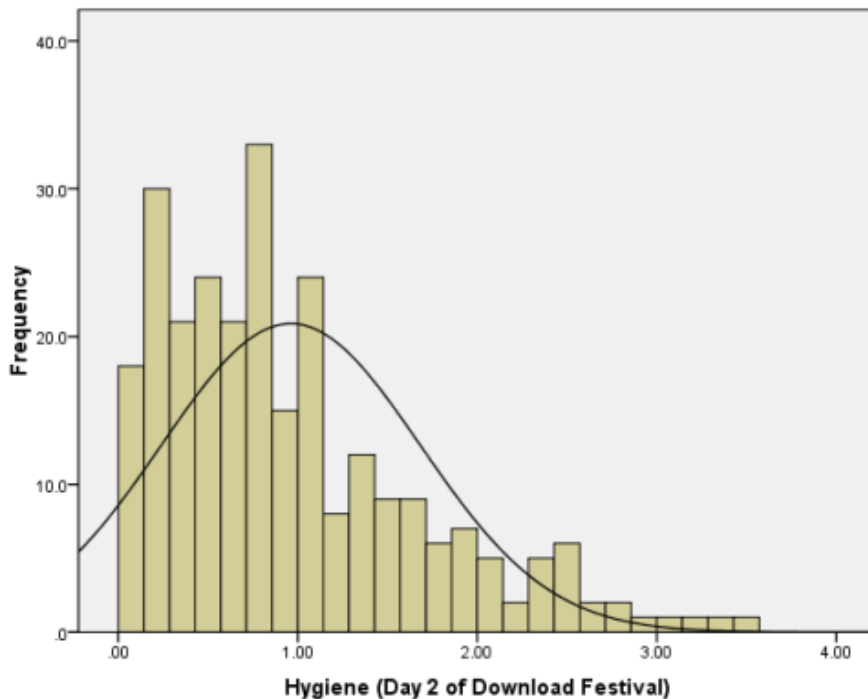
TRANSFORMATION	Reasonable and acceptable domain of x	Range of $f(x)$
$\log(x)$	positive	all

- This transformation is useful when:
 - High degree of variation within samples (e.g., Chl Conc.)
 - Large outliers (tails) and lots of zeros
- Note: to log-transform data containing zeros, a small number should be added to all data points.
 - With count data, add one, so that: $f(x) = \log(0+1) = 0$
 - With density data, add constant smaller than smallest possible sample, so that: $f(x) = \log(0+0.001) = -3$

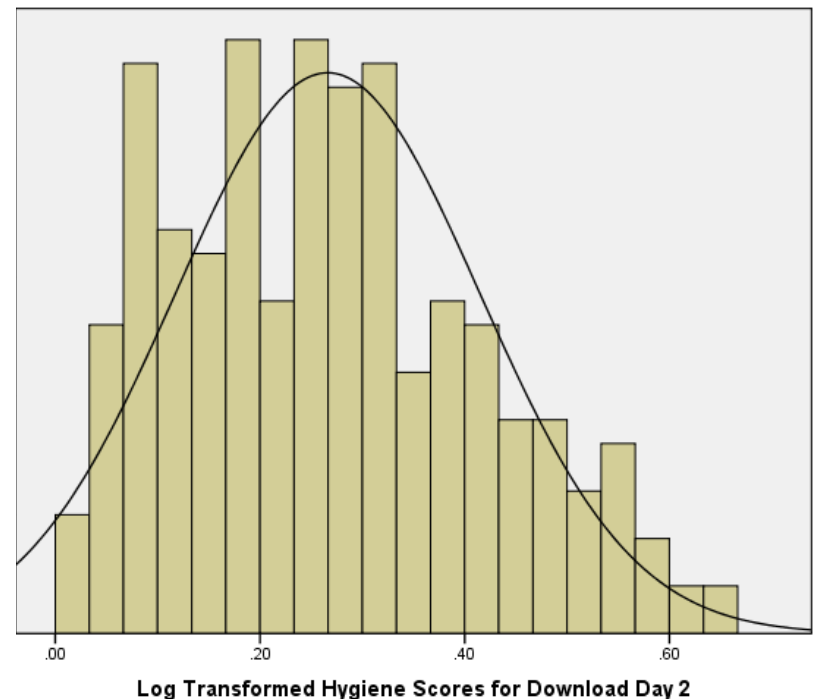
Log Transformation

Log Transformation ($\log(X_i)$): Reduces positive skew

Before



After



Monotonic Data Transformations

Arcsine / Arcsine-squareroot transformation

TRANSFORMATION	Reasonable and acceptable domain of x	Range of $f(x)$
$(2/\pi) \cdot \arcsin(x)$	$0 \leq x \leq 1$	0 to 1 inclusive
$(2/\pi) \cdot \arcsin(x^{1/2})$	$0 \leq x \leq 1$	0 to 1 inclusive

- This transformation is useful when:
 - Dealing with proportional data (e.g., Percent Cover)
- Note: data must range between 0 and 1, inclusive.

The constant $2 / \pi$ scales the result of $\arcsin(x)$ [in radians] to range from 0 to 1, assuming that $0 \leq x \leq 1$.

Approach

Powerful parametric tests are based on assumptions:

- Additionally, parametric tests that compare two or more samples require homogeneity of variances (equal variances across samples)

Can be assessed in three ways:

- Graphically
 - Quantitatively
 - Statistically
- HINT: Having similar sample sizes is a great way to ensure equal variances across samples.

Assessing Variance Homogeneity

Graphs: Scatterplots (e.g., Regressions)

Variance Ratio: (with 2 or more groups)

VR = Largest variance / Smallest variance

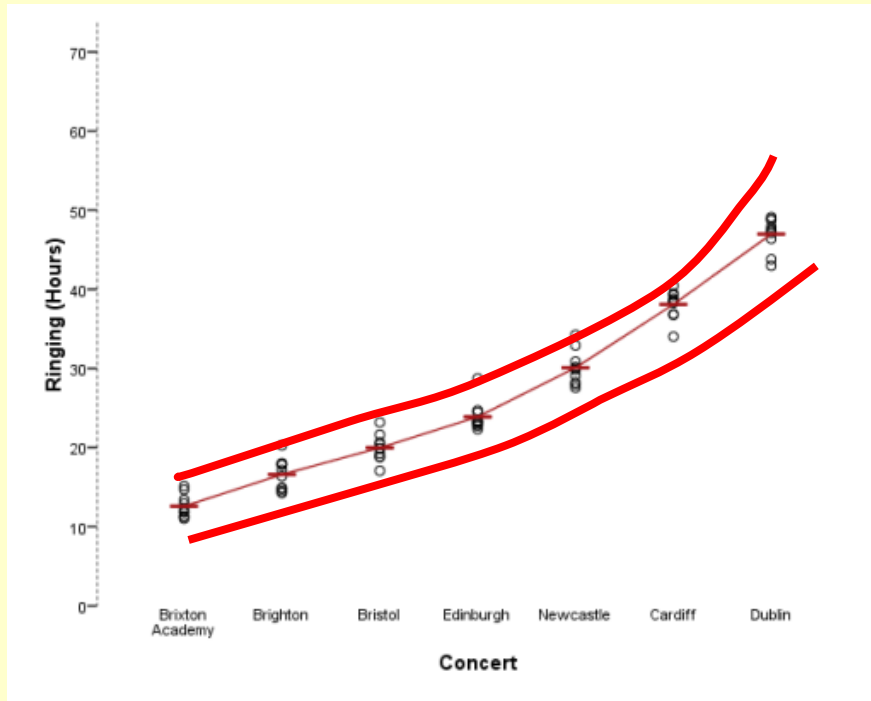
If $VR < 2$, can assume homogeneity

Levene's Test: Tests if variances are equal

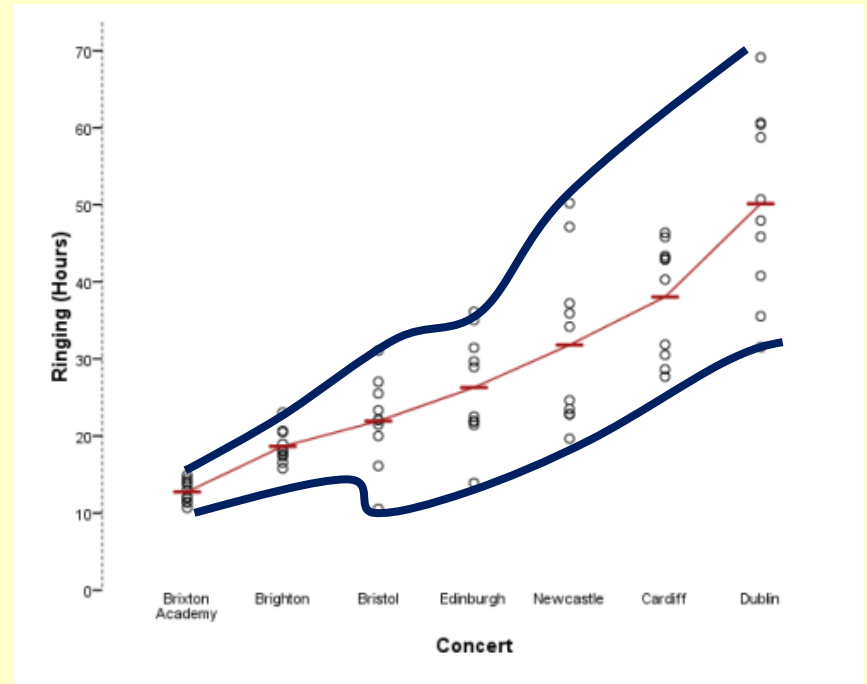
Significant = Variances are not equal

Non-Significant = Variances are equal

Variance Homogeneity - Graphic



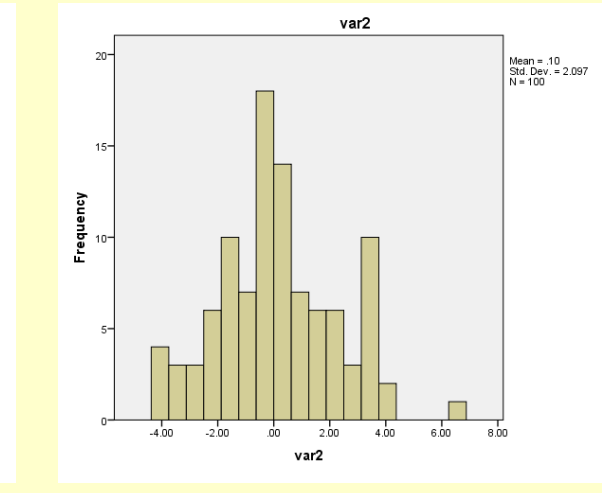
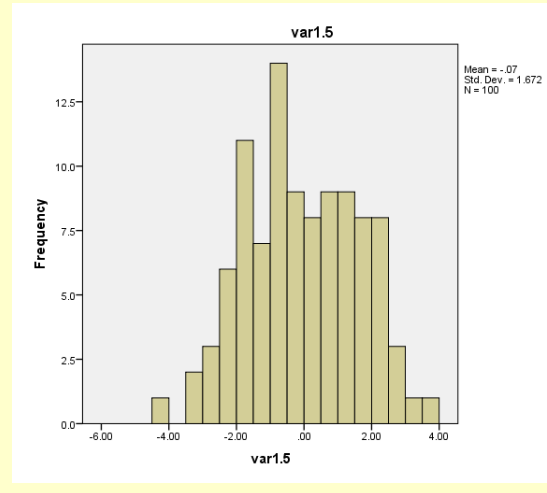
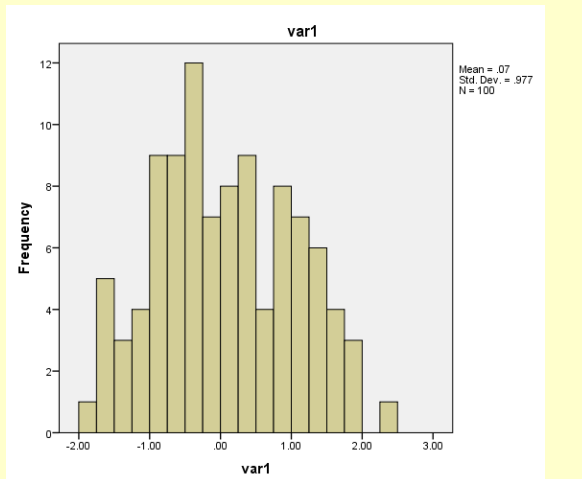
Homogeneous



Heterogeneous

Variance Homogeneity - Ratio

Comparing three normal distributions with the same means and different variances: 1, 2.25, 4



Pairwise Variance Comparisons

<u>Larger</u>	<u>Smaller</u>	<u>Ratio</u>
4	1	4

Rule of Thumb:
Ratio > 2



Variance Homogeneity - Test

data	level
-.71	1.00
-.49	1.00
1.25	1.00
1.57	1.00
-.25	1.00
.68	1.00
-1.24	1.00
-.91	1.00
.78	1.00
1.46	1.00
-.25	1.00
1.19	1.00
1.28	1.00
-.83	1.00
.53	1.00
-1.47	1.00
-.80	1.00

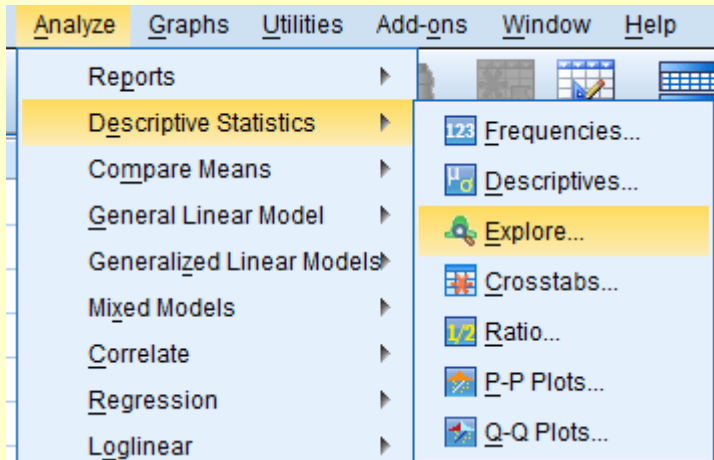
We can test all three distributions at once:

-Use one column for the data
(combine all data in one column)

-Use another column for levels
(of categorical factor: distribution)

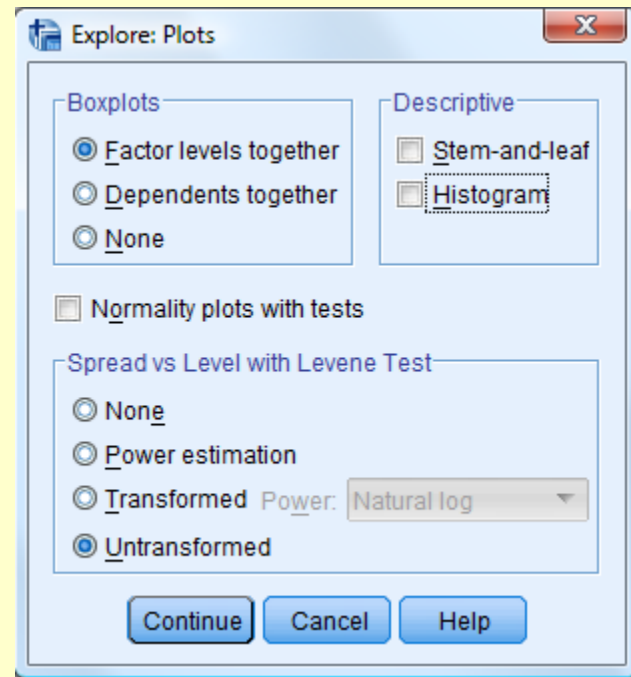
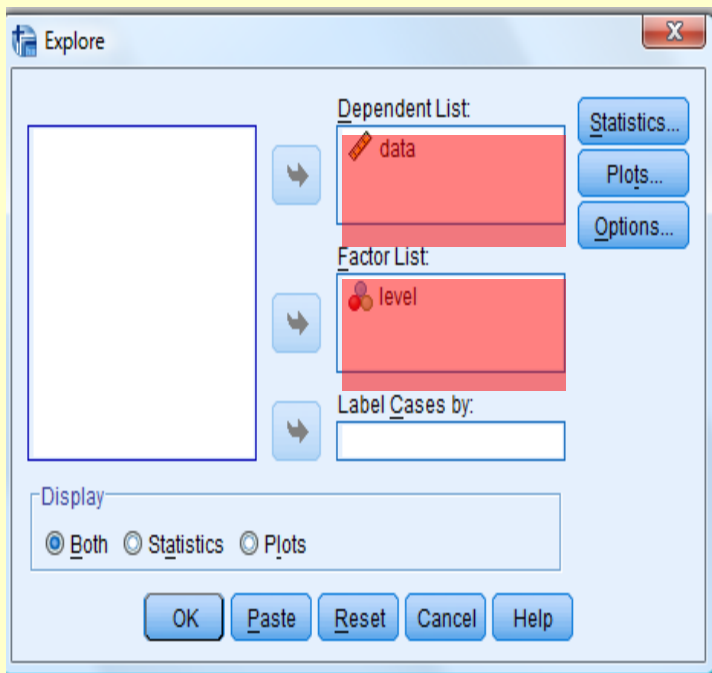
Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
data	Numeric	8	2		None	None	8	≡ Right	 Scale
level	Numeric	8	2		None	None	8	≡ Right	 Nominal

Variance Homogeneity - Test



Identify the dependent list
(the data you are analyzing)

Identify the factor (group)



Boxplots

Levene's
Test
(No
transform)

Variance Homogeneity - Test

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
data	Based on Mean	14.213	2	297	.000
	Based on Median	14.242	2	297	.000
	Based on Median and with adjusted df	14.242	2	255.942	.000
	Based on trimmed mean	14.203	2	297	.000

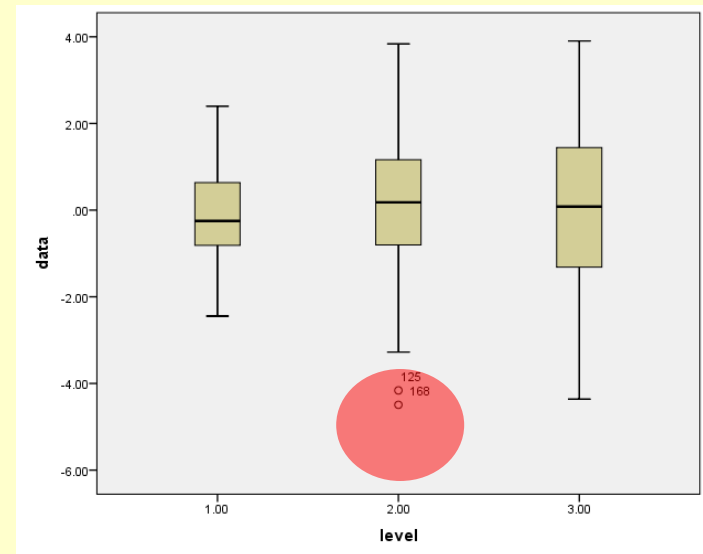
Df = 2 (k-1), 297
(Total = 300 - 1)

p < 0.001
Significant
(Different)

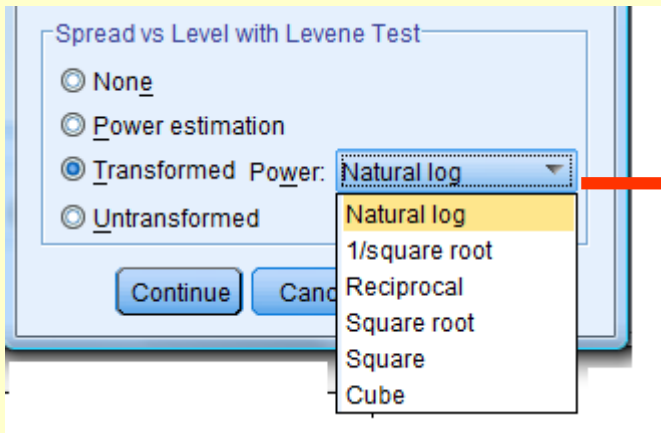
Box Plots:

For each distribution
5, 25, 50, 75, 95 %

Outliers (beyond)



Correcting Data Problems



Transformations

Most Important Rule: Do not Reverse the Order of the Values (larger remains larger... smaller remains smaller)

Monotonic: change values but retain ranks

Non-monotonic: change values and ranks

(For example: Add random number, Multiply by (-1))

Variance Homogeneity - You Try



data	level
-.71	1.00
-.49	1.00
1.25	1.00
1.57	1.00
-.25	1.00
.68	1.00
-1.24	1.00
-.91	1.00
.78	1.00
1.46	1.00
-.25	1.00
1.19	1.00
1.28	1.00
-.83	1.00
.53	1.00
-1.47	1.00
-.80	1.00

Use dataset posted online:
Statistics_Dataset2.xls

- Uses one column for the data

- Uses another column for the
different samples (levels / groups)

(a categorical variable: distribution)

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
data	Numeric	8	2		None	None	8	≡ Right	 Scale
level	Numeric	8	2		None	None	8	≡ Right	 Nominal

Summary

- Parametric tests based on normal distributions
- If possible, we want to use powerful parametric tests
- While significant differences can be achieved more easily with parametric statistics ...
- Critical to ensure a normal underlying data distribution, when using a parametric statistical model to avoid committing type-I error
- **Remember:**
Even if the underlying distribution is normal, normality requires a large enough sample size