Chapter 4
ANOVA Diagnostics and Remedial Measures

Violations of Assumptions in ANOVA

Because everything does not always go as planned . . .

1. Review of assumptions for oneway ANOVA:
   - All samples are drawn from normally distributed populations
   - All populations have a common variance
   - All samples were drawn independently from each other
   - Within each sample, the observations were sampled randomly and independently of each other
   - Factor effects are additive

- In our data, we need to check that:
  - Each sample appears to come from a population with a normal distribution.
  - All samples come from populations with a common variance.
  - There is a lack of outliers.

- The *F* statistic is relatively robust to violations of normality if:
  - The populations are symmetrical and unimodal.
  - The cell sizes are equal and greater than 10.

  - In general, so long as the sample sizes are equal and large, you just need to check that the samples are symmetrical and homogeneous in shape.

- The *F* statistic is NOT robust to violations of homogeneity of variances:
  - Rule of Thumb: If the ratio of the largest variance to smallest variance is less than 3 and the cell sizes are equal, the F-test will be valid.
  - If the sample sizes are unequal then smaller differences in variances can disrupt the F-test.

  - We must pay much more attention to unequal variances than to non-normality of data.

2. Testing the Normality/Symmetry Assumption

- Testing for normality should be conducted on a cell-by-cell basis

- Tests to examine normality:
  - Side-by-side boxplots and histograms
  - Coefficients of skewness and kurtosis
    - Can conduct t-tests, if desired
  - Statistical tests
    - Shapiro-Wilk test
    - Kolmogorov-Smirnov test

- **Statistical Tests of Normality**

- Kolmogorov-Smirnov (KS) test:
  - A general test to detect departures from any specified distribution.
  - It can be used to check normality, but it tends to be less powerful than tests developed specifically to check normality.
  - Loses power if the mean and variance are not known in advance.
  - A commonly used test for historical reasons, but is no longer very useful to test for departures from normality.

- Shapiro-Wilk (SW) test:
  - Designed specifically to check for departures from normality and is more powerful than (KS test).
  - Mean and variance do not need to be specified in advance.
  - In essence, the SW provides a correlation between the raw data and the values would be expected if the observations followed a normal distribution. The SW statistic tests if this correlation is different from 1.
  - The SW is a relatively powerful test of non-normality and is capable of detecting small departures from normality even with small sample sizes.
  - This test is often too powerful for our purposes. Interpret with caution!

- In SPSS:
```
EXAMINE  VARIABLES=dv BY iv
  /PLOT NPPLOT.
```

  - This syntax give both the KS and SW normality tests. SW test is only (consistently) produced if $n < 50$.

  - For both tests:
    H_0: Data are sampled from a normal distribution
    H_1: Data are NOT sampled from a normal distribution

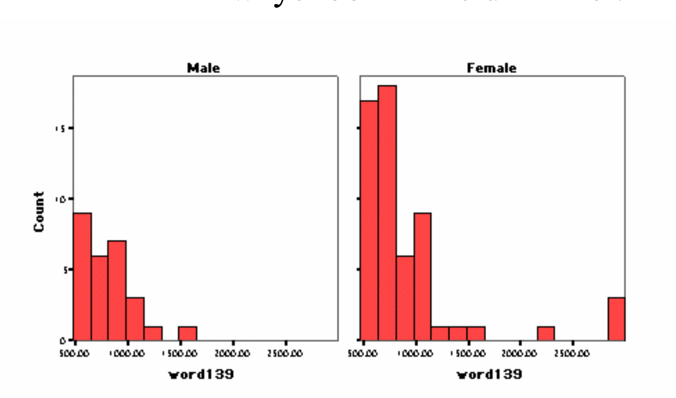
    Rejecting the null hypothesis indicates that the data are non-normally distributed.

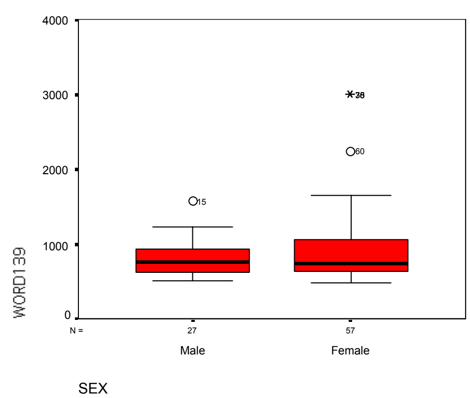
- Example with real data #1: Reaction time responses:
  - Data are reaction times in milliseconds.

  - Are reaction times normally distributed for men and women?

Males	$n = 27$
Females	$n = 57$

  - Always look at the data first!





- Then you can look at the statistics and tests:

Tests of Normality

	SEX	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
WORD139	1.00	.120	27	.200*	.904	27	.017
	2.00	.232	57	.000	.645	57	.000

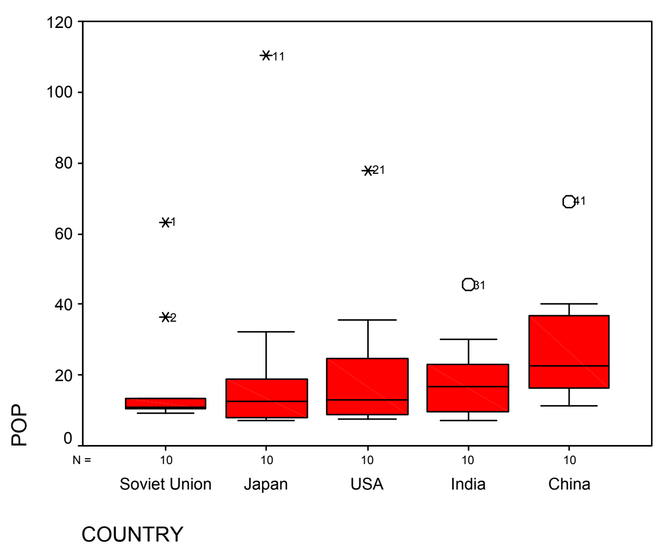
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

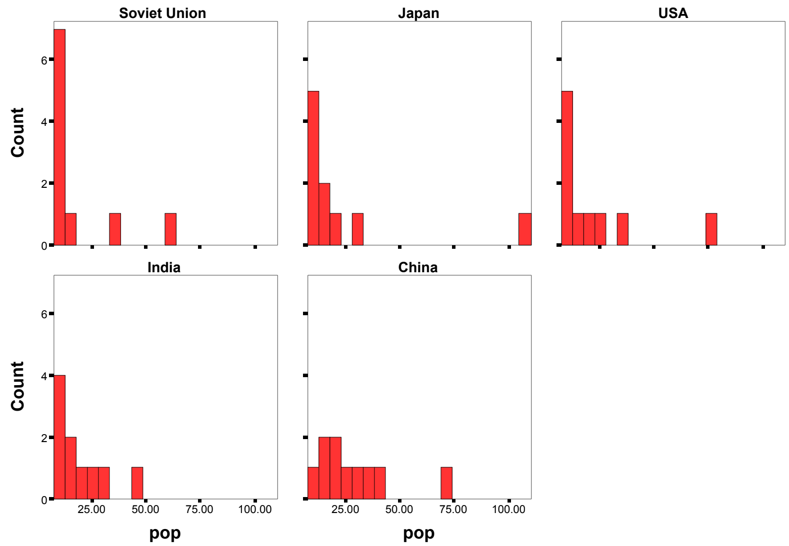
Descriptives

	SEX		Statistic	Std. Error
WORD139	Male	Mean	813.7037	46.24929
		Median	753.0000	
		Variance	57752.909	
		Std. Deviation	240.31835	
		Range	1077.00	
		Interquartile Range	310.0000	
		Skewness	1.311	.448
		Kurtosis	2.602	.872
	Female	Mean	939.3509	76.91656
		Median	737.0000	
		Variance	337220.9	
		Std. Deviation	580.70728	
		Range	2515.00	
		Interquartile Range	432.0000	
		Skewness	2.638	.316
		Kurtosis	6.937	.623

- Example with real data #2: Population of the 10 largest cities of the 16 largest countries (in 1960):
  - Population is given in 100,000s.
  - For the sake of presentation, let's focus on the 5 largest countries.



- Are the populations of the 10 largest cities normally distributed for all five countries?



Tests of Normality

	COUNTRY	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
POP	Soviet Union	.417	10	.000	.586	10	.000
	Japan	.360	10	.001	.560	10	.000
	USA	.256	10	.062	.701	10	.001
	India	.166	10	.200*	.876	10	.118
	China	.208	10	.200*	.857	10	.071

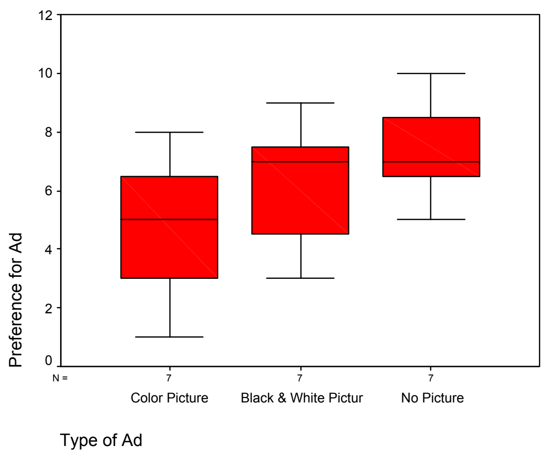
*. This is a lower bound of the true significance.

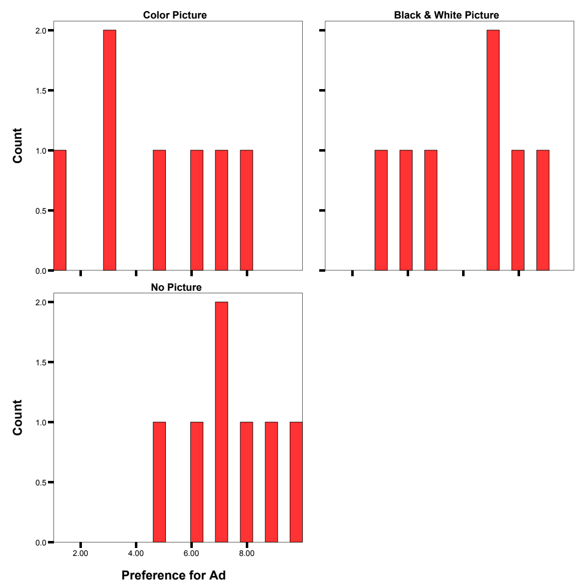
a. Lilliefors Significance Correction

Descriptives

	COUNTRY		Statistic	Std. Error
POP	Soviet Union	Mean	18.5770	5.59789
		Median	10.8700	
		Skewness	2.284	.687
		Kurtosis	4.882	1.334
	Japan	Mean	23.6280	9.90443
		Median	12.6600	
		Skewness	2.856	.687
		Kurtosis	8.467	1.334
	USA	Mean	21.7480	6.86900
		Median	13.0450	
		Skewness	2.263	.687
		Kurtosis	5.534	1.334
	India	Mean	18.9600	3.69945
		Median	16.6800	
		Skewness	1.384	.687
		Kurtosis	1.973	1.334
	China	Mean	28.7630	5.38377
		Median	22.7850	
		Skewness	1.585	.687
		Kurtosis	2.945	1.334

- Example with real data #3: An Advertising Example
    - Three conditions:
        - Color picture n=7
        - Black and white picture n=7
        - No picture n=7
    - Are the favorability ratings normally distributed for all three conditions?





Tests of Normality

	Type of Ad	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Preference for Ad	Color Picture	.182	7	.200*	.961	7	.827
	Black & White Picture	.223	7	.200*	.949	7	.720
	No Picture	.170	7	.200*	.980	7	.958

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Descriptives

	Type of Ad		Statistic	Std. Error
Preference for Ad	Color Picture	Mean	4.7143	.94401
		Median	5.0000	
		Std. Deviation	2.49762	
		Interquartile Range	4.0000	
		Skewness	-.176	.794
		Kurtosis	-1.152	1.587
	Black & White Picture	Mean	6.1429	.82890
		Median	7.0000	
		Std. Deviation	2.19306	
		Interquartile Range	4.0000	
		Skewness	-.252	.794
		Kurtosis	-1.366	1.587
	No Picture	Mean	7.4286	.64944
		Median	7.0000	
		Std. Deviation	1.71825	
		Interquartile Range	3.0000	
		Skewness	.169	.794
		Kurtosis	-.638	1.587

- A final word on checking normality:
  - Remember that normality is the least important of the ANOVA assumptions.
  - Large samples and equal cell sizes make life much easier.
  - So long as all cells show the same distribution of data (and cell sizes are relatively equal) and are not excessively deviant, no remedial measures are necessary.

3. Testing the Equality of Variances Assumption

- When we derived the F-test, we assumed that the variances in each condition were identical.
  - F-test is NOT robust to violations of homogeneity of variance.
  - We need to be more watchful for violation of the equality of variances assumption than we were for the normality assumption.

- Tests to examine homogeneity of variances:
  - Side-by-side boxplots
  - Variance/Standard Deviation/IQR statistics
  - Levine's Test

- Levene's test of homogeneity of variances:
  - For Levene's test, the residuals from the cell means are calculated:

    For group j: $e_{ij} = Y_{ij} - \overline{Y}_j$

  - An ANOVA is then conducted on the absolute value of the residuals. If the variances are equal in all groups, then the average size of the residual should be the same across all groups.

  - For Levene's test, we have the following null and alternative hypotheses:

    H_0: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_a^2$

    H_1: Not all variances are equal

    - Heterogeneity of variances is suggested when you reject the null hypothesis.

  - An example:
    - Raw Data

Group 1	Group 2	Group 3
5	6	4
5	7	7
3	5	2
4	6	8
3	6	9
$\overline{X}_1 = 4$	$\overline{X}_2 = 6$	$\overline{X}_3 = 6$
$s_1^2 = 1$	$s_2^2 = 0.5$	$s_3^2 = 8.5$

    - Take the Absolute Value of the Residuals:

Group 1	Group 2	Group 3
1	0	2
1	1	1
1	1	4
0	0	2
1	0	3

- Conduct an ANOVA on the absolute value of the residuals:

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	11.2	2	5.6	9.333333	0.00359	3.88529
Within Groups	7.2	12	0.6			
Total	18.4	14				

- Or you can obtain Levene's test directly from SPSS:

```
EXAMINE VARIABLES=dv BY group
 /PLOT spreadlevel.
```

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
DV	Based on Mean	9.333	2	12	.004
	Based on Median	3.190	2	12	.077
	Based on Median and with adjusted df	3.190	2	5.106	.126
	Based on trimmed mean	8.876	2	12	.004

- From our hand calculations: $F(2,12) = 9.33, p < .01$
- From SPSS (based on mean): $F(2,12) = 9.33, p < .01$

- Variations on Levene's test:
  - Based on the median

    For group j: $e'_{ij} = Y_{ij} - Median_j$

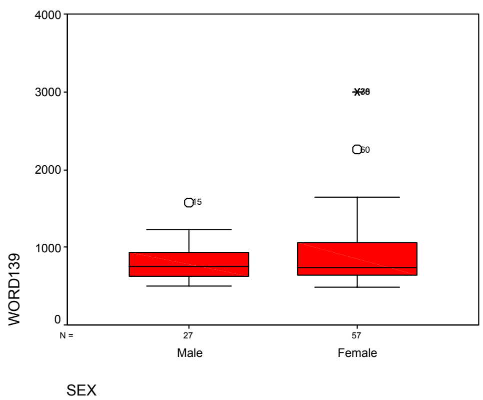
  - Based on trimmed mean

    First toss out 5% of the largest observations and 5% of the smallest observations. Then calculate the mean and proceed as usual.

- Words of caution about Levene's test:
  - Need to assume that the absolute value of the residuals satisfy the assumptions of ANOVA.
  - Most people use a more liberal cut off value when testing homogeneity of variances (due to the poor power of these tests).

- Example with real data #1: Reaction time responses
  - Do the reaction times have equal variances for men and women?

    Males $n = 27$ Females $n = 57$



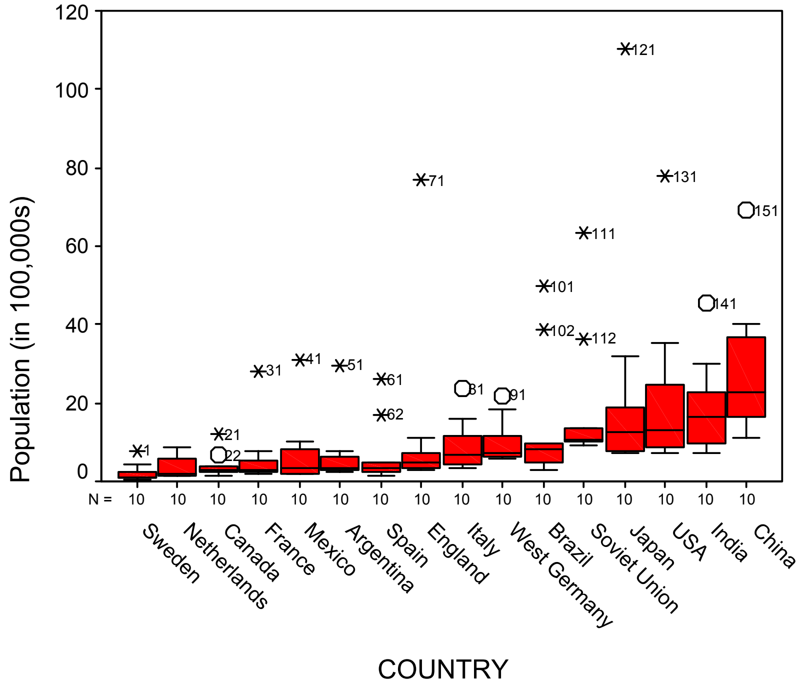
Descriptives

	SEX		Statistic	Std. Error
WORD139	Male	Mean	813.7037	46.24929
		Variance	57752.909	
		Std. Deviation	240.31835	
		Minimum	501.00	
		Maximum	1578.00	
		Range	1077.00	
		Interquartile Range	310.0000	
	Female	Mean	939.3509	76.91656
		Variance	337220.9	
		Std. Deviation	580.70728	
		Minimum	485.00	
		Maximum	3000.00	
		Range	2515.00	
		Interquartile Range	432.0000	

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
WORD139	Based on Mean	4.317	1	82	.041
	Based on Median	1.971	1	82	.164
	Based on Median and with adjusted df	1.971	1	61.202	.165
	Based on trimmed mean	2.908	1	82	.092

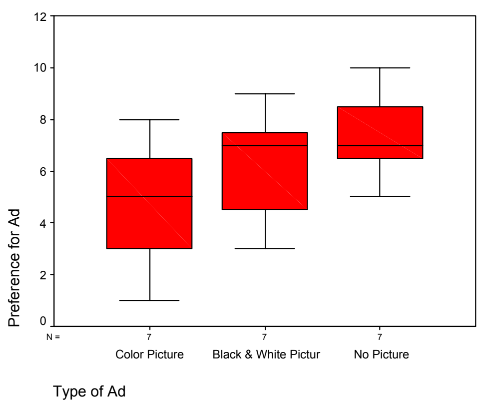
- Example with real data #2: Population of the 10 largest cities of the 16 largest countries (in 1960)
  - Are the variances of the 10 largest cities equal for all 16 countries?



Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
POP	Based on Mean	2.465	15	144	.003
	Based on Median	.992	15	144	.467
	Based on Median and with adjusted df	.992	15	53.533	.476
	Based on trimmed mean	1.690	15	144	.059

- Example with real data #3: An Advertising Example
    - Three conditions:
        - Color picture
        - Black and white picture
        - No picture
    - Are the variances of the favorability ratings equal for all three conditions?



Descriptives

	Type of Ad		Statistic	Std. Error
Preference for Ad	Color Picture	Mean	4.7143	.94401
		Median	5.0000	
		Variance	6.238	
		Std. Deviation	2.49762	
		Minimum	1.00	
		Maximum	8.00	
		Range	7.00	
		Interquartile Range	4.0000	
		Skewness	-.176	.794
		Kurtosis	-1.152	1.587
	Black & White Picture	Mean	6.1429	.82890
		Median	7.0000	
		Variance	4.810	
		Std. Deviation	2.19306	
		Minimum	3.00	
		Maximum	9.00	
		Range	6.00	
		Interquartile Range	4.0000	
		Skewness	-.252	.794
		Kurtosis	-1.366	1.587
	No Picture	Mean	7.4286	.64944
		Median	7.0000	
		Variance	2.952	
		Std. Deviation	1.71825	
		Minimum	5.00	
		Maximum	10.00	
		Range	5.00	
		Interquartile Range	3.0000	
		Skewness	.169	.794
		Kurtosis	-.638	1.587

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Preference for Ad	Based on Mean	.865	2	18	.438
	Based on Median	.528	2	18	.599
	Based on Median and with adjusted df	.528	2	17.028	.599
	Based on trimmed mean	.851	2	18	.443

4. Testing for outliers

- Tests to examine outliers:
  - o Side-by-side boxplots and histograms of the raw data
  - o Examine the residuals:
    - Look at standardized residuals
    - Plot of residuals by group

- Examining residuals:

For group j: $e_{ij} = Y_{ij} - \bar{Y}_j$

  - o The residual is a measure of how far away an observation is from its predicted value (our best guess of the value).
  - o If an observation has a large residual, we consider it an outlier.
  - o How large is large? We usually think in terms of standard deviations from the mean, so it would be convenient to standardize the residuals.

- Standardized residual defined:
  - o Recall that for a $N(\mu, \sigma)$ variable, a z-score is computed by:

$$z = \frac{Y_{obs} - \mu}{\sigma}$$

    - For one way ANOVA, the observed residual is equal to:

$$e_{ij} = Y_{ij} - \bar{Y}_j$$

    - And if the population is normally distributed, then the residuals are also normally distributed: $\varepsilon \sim N(0, \sqrt{MSW})$

$$\tilde{e}_{ij} = \frac{e_{ij} - 0}{\sqrt{MSW}} = \frac{Y_{ij} - \bar{Y}_{\cdot j}}{\sqrt{MSW}}$$

  - o Standardized residuals can be interpreted as z-scores.
  - o If the data are normally distributed, then $\tilde{\varepsilon} \sim N(0,1)$ and
    - About 5% of the observations are expected to have a $|\tilde{\varepsilon}| > 2|$
    - About 1% of the observations are expected to have a $|\tilde{\varepsilon}| > 2.5$
  - o For modest sample sizes, $|\tilde{\varepsilon}| > 2.5$ is a reasonable cutoff to call a point an outlier.

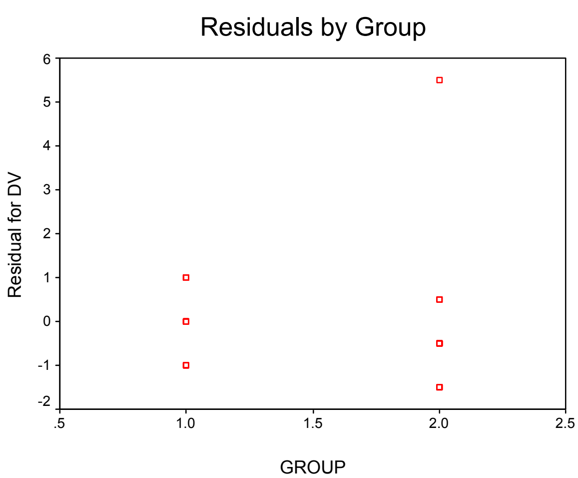
  - o Standardized and Unstandardized residuals give you the same information; it is just a matter of which you prefer to examine.

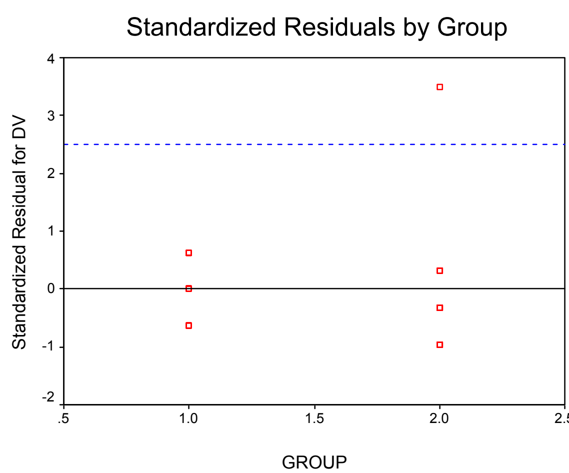
Raw Data		Residuals		Z-Residuals	
Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
3	4	-1	-1.5	-0.64	-0.95
4	5	0	-0.5	0.00	-0.32
5	6	1	0.5	0.64	0.32
4	5	0	-0.5	0.00	-0.32
3	4	-1	-1.5	-0.64	-0.95
4	5	0	-0.5	0.00	-0.32
5	6	1	0.5	0.64	0.32
4	5	0	-0.5	0.00	-0.32
3	4	-1	-1.5	-0.64	-0.95
5	11	1	5.5	0.64	3.50
$\overline{X}_1 = 4$	$\overline{X}_1 = 5.5$				
$\sqrt{MSE} = 1.5723$					

- To calculate residuals in SPSS:

```
UNIANOVA  dv  BY iv
   /SAVE = RESID ZRESID.

UNIANOVA  dv  BY iv
  /SAVE = RESID (chubby) ZRESID (flubby).
```





- Example with real data #1: Reaction time responses
  - Are there any outliers?

    Males $n = 27$ Females $n = 57$

  - First, look for large outliers:

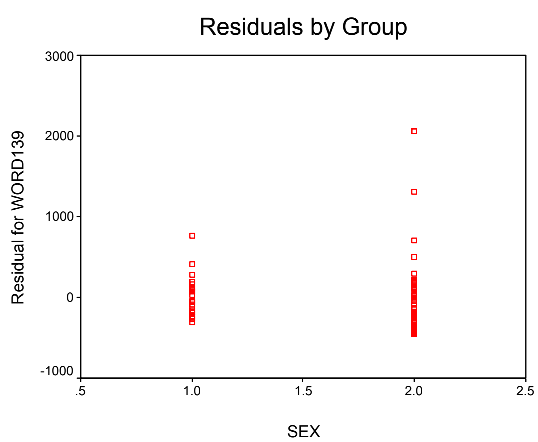
```
UNIANOVA  word139  BY sex
 /SAVE = RESID (resid) ZRESID (zresid).
EXAMINE  VARIABLES=resid BY sex
 /STAT=EXTREME.
```

Extreme Values

	SEX			Case Number	Value
Residual for WORD139	Male	Highest	1	15	764.30
			2	71	416.30
			3	43	274.30
			4	22	190.30
			5	66	185.30
		Lowest	1	79	-312.70
			2	41	-269.70
			3	46	-263.70
			4	25	-238.70
			5	11	-237.70
	Female	Highest	1	35	2060.65
			2	78	2060.65
			3	30	2060.65
			4	60	1313.65
			5	13	705.65
		Lowest	1	45	-454.35
			2	27	-440.35
			3	53	-423.35
			4	28	-419.35
			5	9	-416.35

  - Next, plot the outliers:

```
GRAPH  /SCATTERPLOT=sex WITH resid
 /TITLE= 'Residuals by Group'.
```



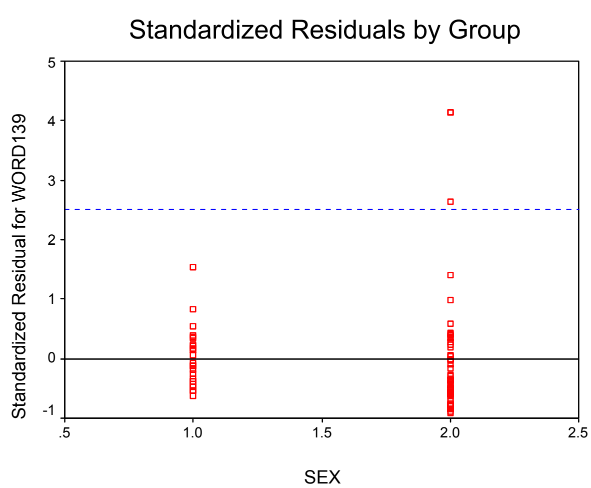
- Or if you prefer, use standardized residuals:

```
EXAMINE VARIABLES=zresid BY sex
        /STAT=EXTREME.
```

Extreme Values

SEX				Case Number	Value
Standardized Residual for WORD139	Male	Highest	1	15	1.53
			2	71	.83
			3	43	.55
			4	22	.38
			5	66	.37
		Lowest	1	79	-.63
			2	41	-.54
			3	46	-.53
			4	25	-.48
			5	11	-.48
	Female	Highest	1	35	4.13
			2	30	4.13
			3	78	4.13
			4	60	2.63
			5	13	1.42
		Lowest	1	45	-.91
			2	27	-.88
			3	53	-.85
			4	28	-.84
			5	9	-.84

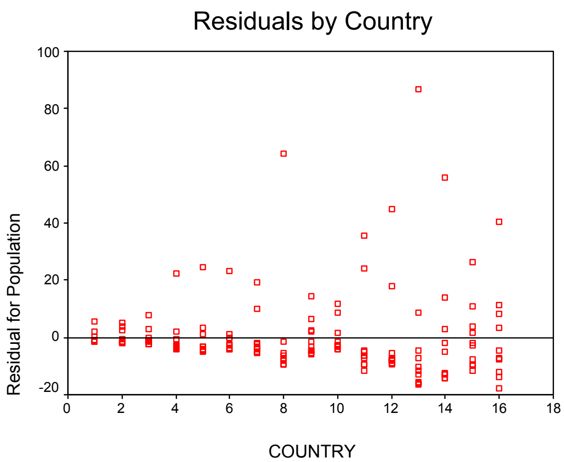
```
GRAPH /SCATTERPLOT=sex WITH zresid
  /TITLE= 'Residuals by Group'.
```



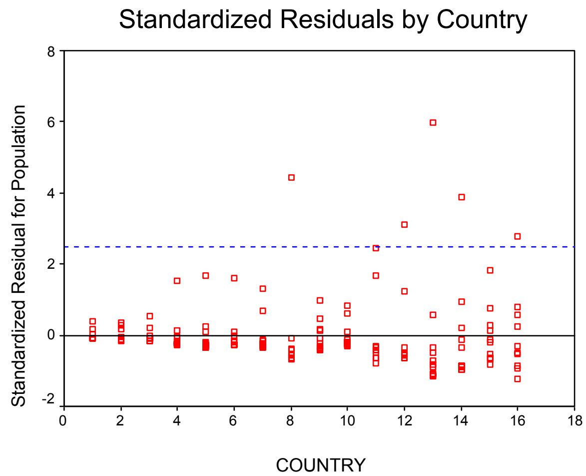
- Example with real data #2: Population of the 10 largest cities of the 16 largest countries (in 1960)
  - Are any of the city populations considered outliers? ($s = 15.84$)

```
UNIANOVA pop  BY country
 /SAVE = ZRESID(zres).
```

```
GRAPH /SCATTERPLOT=country WITH resid
 /TITLE= 'Residuals by Country'.
```




```
GRAPH /SCATTERPLOT=country WITH zresid
  /TITLE= 'Standardized Residuals by Country'.
```



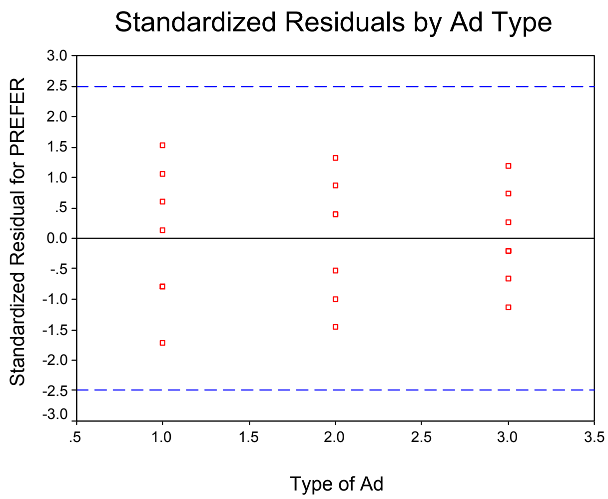
  - You can look at the large residuals to identify them.

```
EXAMINE VARIABLES=zresid BY sex
 /STAT=EXTREME.
```

Extreme Values

	COUNTRY			Case Number	Value
Standardized Residual for POP	USA	Highest	1	131	3.88
			2	132	.95
			3	133	.21
			4	134	-.12
			5	135	-.35
		Lowest	1	140	-.99
			2	139	-.98
			3	138	-.90
			4	137	-.86
			5	136	-.85
	China	Highest	1	151	2.78
			2	152	.78
			3	153	.56
			4	154	.24
			5	155	-.32
		Lowest	1	160	-1.22
			2	159	-.95
			3	158	-.85
			4	157	-.52
			5	156	-.50

- Example with real data #3: An Advertising Example
  - Three conditions:
    - Color picture
    - Black and white picture
    - No picture
  - Are there any outliers in any of the three conditions?




```
EXAMINE VARIABLES=zresid BY ad
    /STAT=EXTREME.
```

Extreme Values

	Type of Ad			Case Number	Value
Standardized Residual for PREFER	Color Picture	Highest	1	5	1.52
			2	3	1.06
		Lowest	1	6	-1.72
			2	1	-.79
	Black & White Picture	Highest	1	12	1.32
			2	13	.86
		Lowest	1	11	-1.45
			2	8	-.99
	No Picture	Highest	1	15	1.19
			2	19	.73
		Lowest	1	18	-1.12
			2	21	-.66

OK, we have identified any problematic non-normality, heterogeneity, and/or outliers. Now what do we do?

5. Sensitivity Analysis
   - Suppose you identified one or more outliers.
     - Always check your data to make sure the outlier is not a data entry / data coding error.
   - You can conduct a sensitivity analysis to see how much the outlying observations affect your results.

   - How to do a sensitivity analysis:
     - Run an ANOVA on the entire data.
     - Remove outlier(s) and rerun the ANOVA.

     - <u>If the results are the same</u> then you can report the analysis on the full data and report that the outliers did not influence the results.
     - <u>If the results are different</u>, then life is more difficult . . .

   - Example with real data #1: Reaction time responses
     - Data are reaction times in milliseconds.
     - We applied a log transformation to the data, but there are three female outliers.
     - Let's run an ANOVA on the log-transformed data with and without those outliers.

Extreme Values

	SEX			Case Number	Value
Standardized Residual for LN139	Male	Highest	1	15	1.78
			2	71	1.14
			3	43	.83
			4	22	.63
			5	66	.62
		Lowest	1	79	-1.13
			2	41	-.93
			3	46	-.90
			4	25	-.79
			5	11	-.78
	Female	Highest	1	35	3.24
			2	78	3.24
			3	30	3.24
			4	60	2.51
			5	13	1.72
		Lowest	1	45	-1.38
			2	27	-1.31
			3	53	-1.22
			4	28	-1.20
			5	9	-1.19

- First, let's do the analysis with the outliers:

```
ONEWAY ln139 BY sex
 /STAT = desc.
```

Descriptives

LN139

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
Male	27	6.6640	.27404	.05274	6.5556	6.7724
Female	57	6.7288	.43894	.05814	6.6123	6.8452
Total	84	6.7079	.39299	.04288	6.6227	6.7932

ANOVA

LN139

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.077	1	.077	.495	.484
Within Groups	12.742	82	.155		
Total	12.819	83			

- Next, let's remove the outliers and re-do the analysis:

```
temporary.
select if zre_1 < 3.
ONEWAY ln139 BY sex
 /STAT = desc.
```

Descriptives

LN139

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
Male	27	6.6640	.27404	.05274	6.5556	6.7724
Female	54	6.6578	.32565	.04432	6.5689	6.7467
Total	81	6.6599	.30769	.03419	6.5918	6.7279

ANOVA

LN139

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.001	1	.001	.007	.933
Within Groups	7.573	79	.096		
Total	7.574	80			

- Both analyses give the same results. There is no evidence that the outliers influence our conclusions. Thus, we can be confident when we report the analysis of the complete data.

  With outliers: $F(1,82) = 0.50, p = .48$

  Without outliers: $F(1,79) = 0.01, p = .93$

- Example with real data #2: 10 largest city data
  - o We found that a log-transformation stabilized the variances, for the most part.
  - o There are still quite a few outliers.

Extreme Values

	COUNTRY		Case Number	Value
Standardized Residual for LNPOP	Sweden	Highest	1	2.31
		Lowest	10	-.87
	Netherlands	Highest	11	1.62
		Lowest	20	-.89
	Canada	Highest	21	1.72
		Lowest	30	-.85
	France	Highest	31	2.62
		Lowest	40	-.86
	Mexico	Highest	41	2.60
		Lowest	50	-1.10
	Argentina	Highest	51	2.50
		Lowest	60	-.77
	Spain	Highest	61	2.29
		Lowest	70	-1.30
	England	Highest	71	3.25
		Lowest	80	-1.01
	Italy	Highest	81	1.51
		Lowest	90	-1.05
	West Germany	Highest	91	1.19
		Lowest	100	-.59
	Brazil	Highest	101	2.19
		Lowest	110	-1.60
	Soviet Union	Highest	111	1.94
		Lowest	120	-.60
	Japan	Highest	121	2.58
		Lowest	130	-1.04
	USA	Highest	131	2.09
		Lowest	140	-.98
	India	Highest	141	1.35
		Lowest	150	-1.07
	China	Highest	151	1.33
		Lowest	160	-1.07

- First, we conduct the analysis on the full data:

```
ONEWAY lnpop BY country
 /STAT=desc.
```

ANOVA

LNPOP

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	96.819	15	6.455	11.127	.000
Within Groups	83.532	144	.580		
Total	180.350	159			

- Next, we conduct the analysis without the outliers:

```
temporary.
SELECT IF zres < 2.49.  * Eliminate 6 observations *
ONEWAY lnpop BY country
 /STAT=desc.
```

ANOVA

LNPOP

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	99.991	15	6.666	15.539	.000
Within Groups	59.629	139	.429		
Total	159.619	154			

  - It would appear that the outliers do not affect the conclusions you would draw from this data.
  - But be **very careful**. If you run follow-up tests, you need to perform a sensitivity analysis for each and every analysis you run!

- What happens if the outlier does affect the conclusions?
  - Try a non-parametric test.
  - Report analysis with and without the outlier (often done in a footnote).

6. Kruskal-Wallis test
   - The multi-group equivalent of the Mann-Whitney U test
   - Data must be at least ordinal scale
   - Often called ANOVA by ranks test

   - Conceptually:
     - Rank all observations in the entire data set.
     - Perform an ANOVA on the rank scores for each group.

   - The Kruskal-Wallis test is a non-parametric test:
     - No assumptions are made about the type of underlying distribution.
     - However, it is assumed that the shape of the distribution is equal for all groups (so a weaker version of homogeneity of variances is still necessary).
     - No population parameters are estimated (no confidence intervals).
     - Can be used for samples that strongly deviate from normality or when there are a small number of disruptive outliers.

   - The test statistic, H, has an approximate chi-square distribution. We need at least 10 observations per group for this approximation to hold.
   - If there are small sample sizes and many ties, a corrected Kruskal-Wallis test should be used (but is beyond the scope of this course).
   - If the assumptions of ANOVA are satisfied, then it is less powerful than ANOVA.

     H_0: The distribution of scores is equal across all groups
     H_1: The distribution of scores is NOT equal across all groups

   - We will skip the computational details and rely on SPSS!
   - No well-established measure of effect size is available for the Kruskal-Wallis test.

- Example #1: Reaction Time Responses

```
NPAR TESTS
 /K-W=word139   BY sex(1 2).
```

Ranks

	SEX	N	Mean Rank
WORD139	Male	27	42.00
	Female	57	42.74
	Total	84	

Test Statistics[a,b]

	WORD139
Chi-Square	.017
df	1
Asymp. Sig.	.897

a. Kruskal Wallis Test

b. Grouping Variable: SEX

$\chi^2(1) = 0.017, p = .897$

- The K-W test is equivalent to an ANOVA performed on the ranked data.

```
RANK VARIABLES=word139.
ONEWAY rword139 BY sex
 /STAT=desc.
```

Descriptives

RANK of WORD139

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1.00	27	42.00000	22.360680	4.303315	33.15441	50.84559
2.00	57	42.73684	25.485308	3.375612	35.97468	49.49900
Total	84	42.50000	24.391881	2.661372	37.20664	47.79336

ANOVA

RANK of WORD139

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	9.947	1	9.947	.017	.898
Within Groups	49372.053	82	602.098		
Total	49382.000	83			

$F(1,82) = 0.017, p = .898$

- The p-values may differ slightly between the two-test because the K-W test uses a *chi-square* approximation, and the ANOVA by ranks uses an *F* approximation. With large samples, these two approximations are nearly identical, as we can see in this example.

- Example #2: 10 Largest City data

```
NPAR TESTS
 /K-W=pop   BY country(1 16).
```

Ranks

	COUNTRY	N	Mean Rank
POP	Sweden	10	19.95
	Netherlands	10	34.95
	Canada	10	44.40
	France	10	50.15
	Mexico	10	58.40
	Argentina	10	57.00
	Spain	10	60.40
	England	10	74.35
	Italy	10	87.50
	West Germany	10	92.60
	Brazil	10	94.60
	Soviet Union	10	118.05
	Japan	10	117.25
	USA	10	117.80
	India	10	122.30
	China	10	138.30
	Total	160	

Test Statistics[a,b]

	POP
Chi-Square	88.892
df	15
Asymp. Sig.	.000

a. Kruskal Wallis Test

b. Grouping Variable: COUNTRY

KW Test: $\chi^2(15) = 88.89, p < .001$

- Example #3: Keppel's Advertising Example

```
NPAR TESTS
 /K-W=prefer   BY group(1 3).
```

Ranks

	Type of Ad	N	Mean Rank
Preference for Ad	Color Picture	7	7.64
	Black & White Picture	7	11.07
	No Picture	7	14.29
	Total	21	

Test Statistics[a,b]

	Preference for Ad
Chi-Square	4.104
df	2
Asymp. Sig.	.129

a. Kruskal Wallis Test

b. Grouping Variable: Type of Ad

$\chi^2(2) = 4.10, p = .13$

- Note that when there are more than two groups, the Kruskal-Wallis test is an omnibus test, and you cannot conclude which means are different.

- A non-parametric median test is also available.
  - Bonett, D. G., & Price, R. M. (2002). Statistical inference for a linear function of medians: Confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods, 7*, 370-383.

  - This test examines differences in medians across different samples.
  - The median test is not included in SPSS.

7. Brown-Forsythe F* test (1974)
    - A test of differences in means that does not make the homogeneity of variances assumption.
    - (For a more detailed discussion of this and other similar tests, see Maxwell & Delaney, 1990.)

    - The numerator of this test is the SSB calculated the usual way.
    - The denominator is corrected to account for unequal variances.

    - The parts of the Brown-Forsythe F^* test:

Numerator = SSB

Denominator = $\sum_j \left[1 - \frac{n_j}{N}\right] s_j^2$

$n_j =$ # of observations in group j

$N =$ Total number of observations

$s_j^2 =$ Sample variance for group j

$$F^* = \frac{SSB}{\sum_j \left[1 - \frac{n_j}{N}\right] s_j^2}$$

F^* no longer follows an F distribution

We can approximate the distribution of F^* with $F(a-1, f)$

Where a = # of groups

$$f = \frac{1}{\sum_j \frac{g_j^2}{(n_j - 1)}} \qquad g_j = \frac{\left[1 - \frac{n_j}{N}\right] s_j^2}{\sum_j \left[1 - \frac{n_j}{N}\right] s_j^2}$$

- With equal n for each group, $F^* = F$, but the denominator degrees of freedom will be different.
- When the assumptions are satisfied, F^* is slightly less powerful than the standard F test, but it is still an unbiased, valid test.
- When variances are unequal F will be biased, especially when the cell sizes are unequal. In this situation, F^* remains unbiased and valid.

- Brown-Forsythe F^* test in SPSS:

```
ONEWAY word139 BY sex
 /STATISTICS BROWNFORSYTHE.
```

Robust Tests of Equality of Means

WORD139

	Statistic[a]	df1	df2	Sig.
Brown-Forsythe	1.960	1	81.007	.165

a. Asymptotically F distributed.

$$F^*(1, 81.01) = 1.96, p = .17$$

- When you have only two groups: $(\text{Welch's } t)^2 = F^*$

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
WORD139	Equal variances assumed	4.317	.041	-1.079	82	.284	-125.6472	116.48738
	Equal variances not assumed			-1.400	81.007	.165	-125.6472	89.75051

$$t(81.01) = -1.40, p = .165$$
$$(-1.40^2) = 1.96$$

$$F^*(1, 81.01) = 1.96, p = .17$$

- Now that SPSS has incorporated the F^* test into the program, it would be nice to see people adopt it more routinely, especially when cell sizes are unequal.

- Welch's W test (1951) also corrects for unequal variances, but is even more computationally intensive than F^* (and it is not clear that it performs any better than F^*).

8. Selecting an appropriate transformation

- Why transform the data?
  - To achieve homogeneity of the variances
  - To achieve normality of the group distributions
  - To obtain additivity of effects (rare)

    Suppose your theory says the relationship between variables is:

    $y = abc$ (a multiplicative relationship)

    This relationship cannot be decomposed as

    $$y_{ijkl} = \mu + \alpha_j + \beta_k + \delta_l + \varepsilon_{ijkl}$$

    - But if you apply a log transformation

    $$\begin{aligned}\ln(y) &= \ln(abc) \\ &= \ln(a) + \ln(b) + \ln(c)\end{aligned}$$

    - Now this relationship of $\ln(y)$ can be decomposed as

    $$\ln(y_{ijkl}) = \mu + \alpha_j + \beta_k + \delta_l + \varepsilon_{ijkl}$$

- Rules of Thumb:
  - Square-root transformation: $y = \sqrt{x}$
    - Sometimes used for count data
    - May be helpful if means are proportional to the variances
  - Logarithmic transformation: $y = \ln(x)$
    - Sometimes used for reaction time data or positively skewed data
    - May be helpful if means are proportional to the standard deviations
  - Reciprocal transformation: $y = 1/x$
    - Sometimes used for reaction time data
    - May be helpful if the square of the means are proportional to the standard deviations

	Original Scores			Transformed scores (Square Root Transformation)		
	a_1	a_2	a_3	a_1	a_2	a_3
	2	6	12	1.41	2.45	3.46
	1	4	6	1.00	2.00	2.45
	5	2	6	2.24	1.41	2.45
	2	4	10	1.41	2.00	3.16
	1	7	6	1.00	2.65	2.45
$\bar{Y}$ =	2.2	4.6	8.0	1.41	2.10	2.79
s =	1.64	1.95	2.83	0.50	0.48	0.48
s^2 =	2.70	3.80	8.00	0.25	0.23	0.24

Means are proportional to variances
Try a square root transformation

Now the variances are approximately equal!

- Kirk's (1995) trick:
  - Examine the ratio of the largest observation to the smallest observation in each group.
  - Apply each transformation to the largest and smallest observations.
  - Select the transformation that minimizes the ratio.

	Treatment Levels			$\frac{Range_{largest}}{Range_{smallest}}$
	a_1	a_2	a_3	
Largest Score (L)	5	7	12	
Smallest Score (S)	1	2	6	
Range	4	5	6	$6/4 = 1.50$
ln(L)	1.609	1.946	2.485	
ln(S)	0.000	0.693	1.792	
Range	1.609	1.253	0.693	1.609/0.693 = 2.23
$\sqrt{L}$	2.236	2.646	3.464	
$\sqrt{S}$	1.000	1.414	2.449	
Range	1.236	1.232	0.974	1.236/.974 = 1.269
1/L	0.200	0.143	0.083	
1/S	1.000	0.500	0.167	
Range	0.800	0.357	0.083	.800/.083=9.648

- Select the Square Root transformation.

- Spread and Level Plot:
    - Spread = Variability
    - Level = Central Tendency

    - Plot the spread (y-axis) by the level (x-axis).
    - Draw a straight line through the points and find its slope, β.
    - Use $p=1-\beta$ to determine transformation of the form:

      $$y = x^p$$

    - Any transformation of the form $y = x^p$ is a member of the family of power transformations:

$p=2$	$y = x^2$	Square transformation
$p=1$	$y = x^1$	No transformation
$p=0.5$	$y = x^{1/2} = \sqrt{x}$	Square root transformation
$p=0$	$y = x^0 = \ln(x)$	Log transformation
$p=-0.5$	$y = x^{-1/2} = 1/\sqrt{x}$	Inverse square root transformation
$p=-1$	$y = x^{-1} = 1/x$	Reciprocal transformation
$p=-2$	$y = x^{-2} = 1/x^2$	Reciprocal square transformation

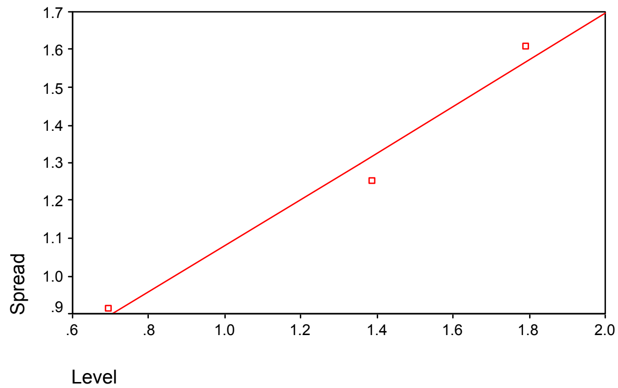
    - In theory, you can use the exact value of p for the transformation, but you may have difficulty explaining and interpreting results based on fractional transformation. It is generally in your best interest to stick with one of these standard options.

    - Which measure of spread and which measure of level?
        - Standard Deviation vs. Mean
        - Standard Deviation vs. Median
        - IQR vs. Median
        - Ln(IQR) vs. Ln(Median) is SPSS's choice

- Spread and level plots in SPSS:

```
EXAMINE VARIABLES=dv BY group
  /PLOT SPREADLEVEL.
```

Spread vs. Level Plot of DV By GROUP



* Plot of LN of Spread vs LN of Level

Slope = .615 Power for transformation = .385

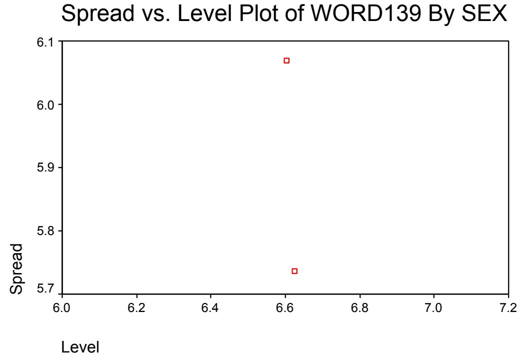
  - From the graph, $p = .385$
  - Round this to the nearest conventional transformation

    $p = .5$ Square root transformation

- Example with real data #1: Reaction time responses
  - Data are reaction times in milliseconds
  - We discovered that the reaction times were positively skewed. Let's try to find a transformation for normality.

  - Let's check the spread and level plot:

Spread vs. Level Plot of WORD139 By SEX



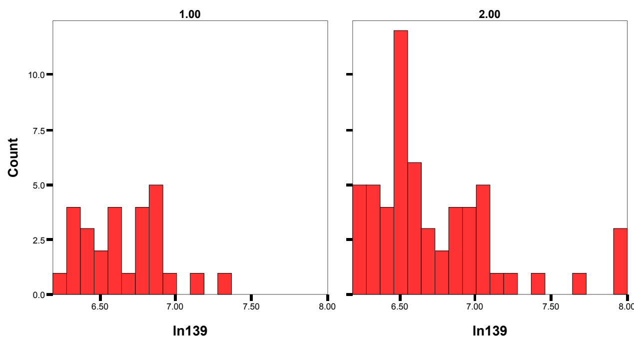
* Plot of LN of Spread vs LN of Level

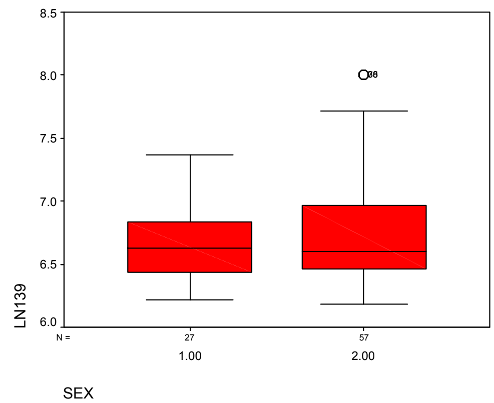
Slope = -15.451 Power for transformation = 16.451

  - Not much help!

- Let's try the rule of thumb that reaction time data should be log transformed.

```
compute ln139 = ln(word139).
```





Descriptives

	SEX			Statistic	Std. Error
LN139	1.00	Mean		6.6640	.05274
		95% Confidence Interval for Mean	Lower Bound	6.5556	
			Upper Bound	6.7724	
		5% Trimmed Mean		6.6524	
		Median		6.6241	
		Variance		.075	
		Std. Deviation		.27404	
		Minimum		6.22	
		Maximum		7.36	
		Range		1.15	
		Interquartile Range		.4028	
		Skewness		.487	.448
		Kurtosis		.113	.872
	2.00	Mean		6.7288	.05814
		95% Confidence Interval for Mean	Lower Bound	6.6123	
			Upper Bound	6.8452	
		5% Trimmed Mean		6.6865	
		Median		6.6026	
		Variance		.193	
		Std. Deviation		.43894	
		Minimum		6.18	
		Maximum		8.01	
		Range		1.82	
		Interquartile Range		.5180	
		Skewness		1.464	.316
		Kurtosis		2.119	.623

- The log transformation appears to fix all problems.

- We can perform an ANOVA on the log-transformed scores.
  ONEWAY ln139 BY sex
  /STAT = ALL.

Descriptives

LN139

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	Upper Bound
1.00	27	6.6640	.27404	.05274	6.5556	6.7724
2.00	57	6.7288	.43894	.05814	6.6123	6.8452
Total	84	6.7079	.39299	.04288	6.6227	6.7932
Model Fixed Effects			.39419	.04301	6.6224	6.7935

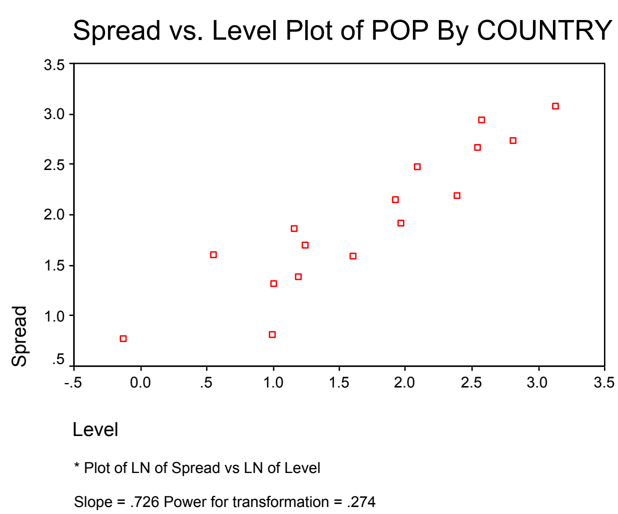
ANOVA

LN139

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.077	1	.077	.495	.484
Within Groups	12.742	82	.155		
Total	12.819	83			

$F(1,82) = 0.50, p = .48, d = .16$

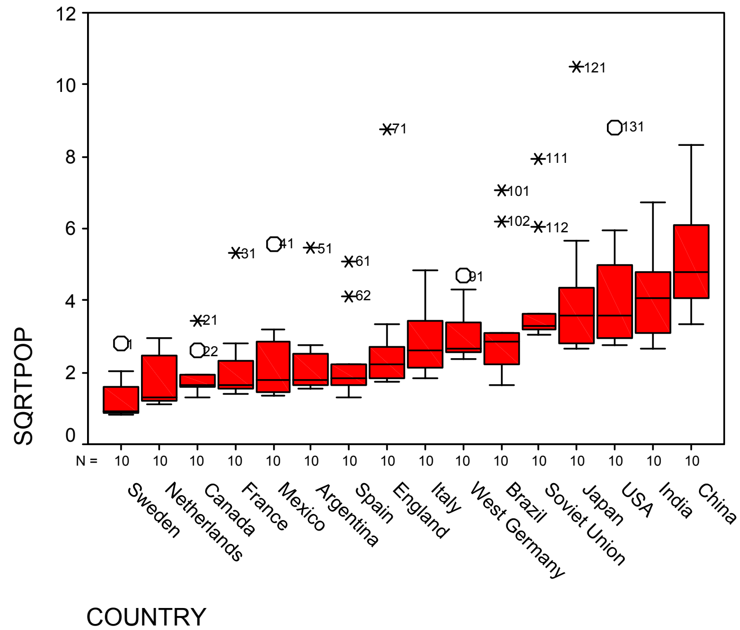
- Example with real data #2:
  - Population of the 10 largest cities of the 16 largest countries (in 1960)
    EXAMINE VARIABLES=pop BY country
    /PLOT SPREADLEVEL.



- The spread and level plot says $p = .274$
- Half way between log transformation and square root transformation; Let's try them both!

○ First, the square root transformation:

compute sqrtpop = sqrt(pop).



Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
SQRTPOP	Based on Mean	1.124	15	144	.340
	Based on Median	.614	15	144	.860
	Based on Median and with adjusted df	.614	15	87.270	.856
	Based on trimmed mean	.857	15	144	.614

- This transformation greatly improved the inequality of the variances

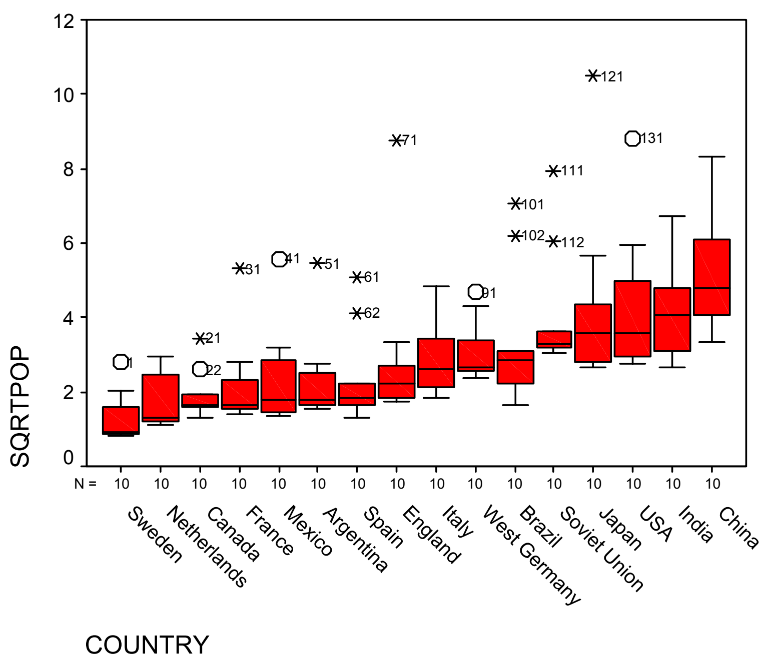
* What does this transformation do for the normality of the data?

Tests of Normality

	COUNTRY	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
SQRTPOP	Sweden	.375	10	.000	.707	10	.001
	Netherlands	.305	10	.009	.772	10	.007
	Canada	.320	10	.005	.770	10	.006
	France	.309	10	.007	.652	10	.000
	Mexico	.308	10	.008	.727	10	.002
	Argentina	.285	10	.021	.646	10	.000
	Spain	.336	10	.002	.750	10	.004
	England	.343	10	.001	.572	10	.000
	Italy	.174	10	.200*	.907	10	.262
	West Germany	.289	10	.018	.780	10	.008
	Brazil	.363	10	.001	.760	10	.005
	Soviet Union	.389	10	.000	.629	10	.000
	Japan	.295	10	.013	.696	10	.001
	USA	.238	10	.115	.806	10	.017
	India	.138	10	.200*	.939	10	.540
	China	.173	10	.200*	.933	10	.483

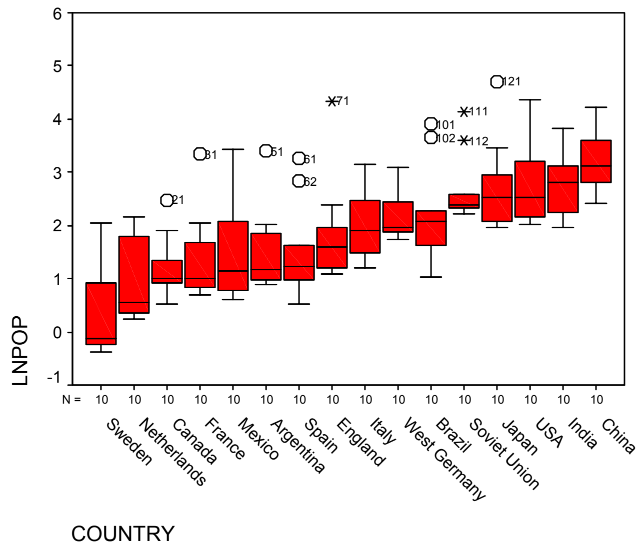
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



* The data from most of the countries still looks skewed and non-normal.

- Now, let's investigate the log transformation:
  compute lnpop = ln(pop).



Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
LNPOP	Based on Mean	.460	15	144	.956
	Based on Median	.260	15	144	.998
	Based on Median and with adjusted df	.260	15	117.390	.998
	Based on trimmed mean	.404	15	144	.976

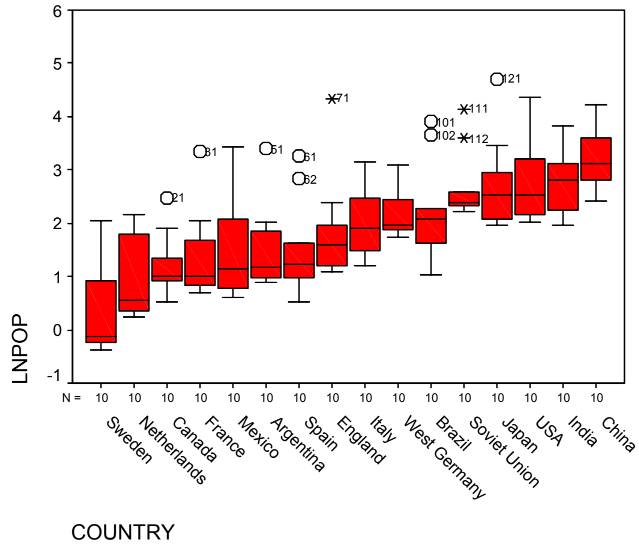
- This does not look bad at all, but what does this transformation do for the normality of the data?

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	COUNTRY	Statistic	df	Sig.	Statistic	df	Sig.
LNPOP	Sweden	.359	10	.001	.756	10	.004
	Netherlands	.288	10	.019	.803	10	.016
	Canada	.290	10	.017	.851	10	.059
	France	.271	10	.036	.777	10	.008
	Mexico	.257	10	.061	.849	10	.056
	Argentina	.236	10	.120	.766	10	.006
	Spain	.255	10	.064	.860	10	.077
	England	.254	10	.066	.750	10	.004
	Italy	.174	10	.200*	.937	10	.519
	West Germany	.258	10	.057	.822	10	.027
	Brazil	.288	10	.018	.875	10	.115
	Soviet Union	.349	10	.001	.676	10	.000
	Japan	.206	10	.200*	.845	10	.051
	USA	.250	10	.076	.880	10	.131
	India	.125	10	.200*	.971	10	.902
	China	.130	10	.200*	.979	10	.962

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



- The log transformation appears to have greatly improved the situation.

  - Now that we have stabilized the variances and the data appear to be roughly normally distributed, we can run an ANOVA on the log-transformed data. However, we will have to make all of our conclusions on the log-transformed scale.

- Example with real data #3: An Advertising Example
  - We determined that each sample was approximately normally distributed, with approximately equal variances and no outliers. Hence, no transformation is necessary.

9. Comparison of Methods for comparing differences between two or more groups
   - Note: All of these tests require
     - Independent groups
     - Within each group, observations must be independent and randomly selected

Method	**When appropriate:**	**Advantages:**	**Disadvantages:**
Parametric tests ANOVA	• Normal/symmetrical data • Equal variances • No outliers	• Most powerful when all assumptions are met • Most familiar	• Gives wrong results when assumptions are not met
Modifications of parametric tests F^*	• Normal/symmetrical data • No outliers	• Requires fewer assumptions • More powerful in typical data	• Less familiar
Transformations	**Transformed** data are: • Normal/symmetrical • Homogeneous in the variances • Without outliers	• Permits use of familiar parametric tests	• May distort meaning of data • Can not always be applied • Conclusions apply to transformed data
Rank-Order Methods K-W Test	• The shape of each distribution must be similar (a weak homogeneity of variances assumption) • $n \geq 10$	• Does not distort data • Can use ordinal data	• Loses information • May be less powerful • Less familiar

10. Examples and Conclusions

- Example #1: Reaction time data
  - What we found:
    - Data have a large positive skew that is similar for both groups
    - Heterogeneity of variances
    - Three outliers, all females
  - What to do:
    - Log transformation with sensitivity analysis
    - Kruskal-Wallis test

Log transformation: $F(1,82) = .50, p = .48$
Log transformation, outliers removed: $F(1,79) = .01, p = .93$
Can report log transformation and footnote results with outliers removed.

Descriptives

LN139

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
Male	27	6.6640	.27404	.05274	6.5556	6.7724
Female	57	6.7288	.43894	.05814	6.6123	6.8452
Total	84	6.7079	.39299	.04288	6.6227	6.7932

ANOVA

LN139

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.077	1	.077	.495	.484
Within Groups	12.742	82	.155		
Total	12.819	83			

Confidence intervals: $\bar{X}_{.j} \pm \left(t_{crit}(df_W) * \sqrt{\frac{MSW}{n_j}} \right)$

For 95% CI: $t_{crit}(82) = 1.99$

Males: $6.664 \pm \left(1.99 * \sqrt{\frac{.155}{27}} \right)$ (6.513, 6.815)

Females: $6.729 \pm \left(1.99 * \sqrt{\frac{.155}{57}} \right)$ (6.625, 6.833)

- Convert CIs back to original scale (for presentation purposes only!)

Males:	$(e^{6.513}, e^{6.815})$	(673.84, 911.41)
Females:	$(e^{6.625}, e^{6.833})$	(753.70, 927.97)

- Effect size

$$\hat{\omega}^2 = \frac{SSBetween - (a-1)MSWithin}{SSTotal + MSWithin}$$

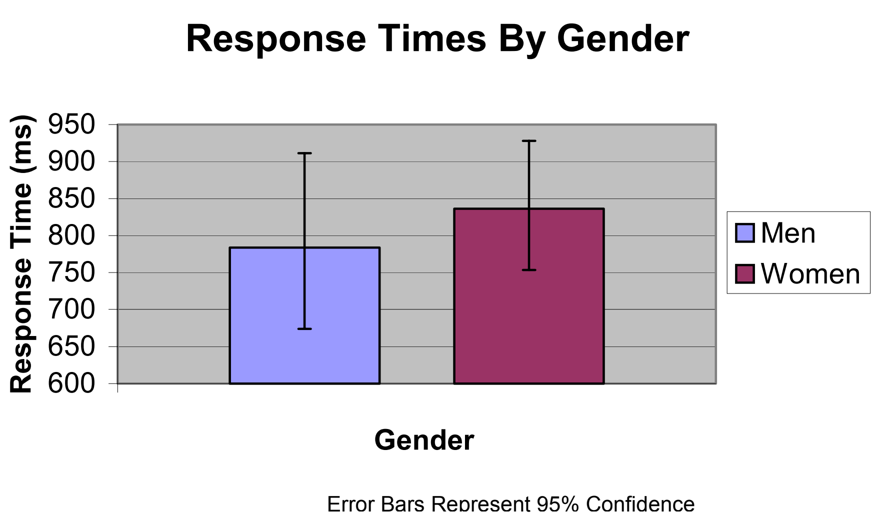
$$\hat{\omega}^2 = \frac{.077 - (1)0.155}{12.819 + .155} = -.006$$

  - Omega squared must be positive.
  - Never report a negative percentage of variance accounted for!! Instead, report $\hat{\omega}^2 < .01$

$$\hat{\sigma}_m = \sqrt{\frac{\sum_{j=1}^{a}(\mu._j - \mu.)^2}{a}}$$

$$\hat{\sigma}_m = \sqrt{\frac{(6.6640 - 6.7079)^2 + (6.7288 - 6.7079)^2}{2}} = .0343$$

$$f = \frac{\hat{\sigma}_m}{\hat{\sigma}_e} = \frac{.0343}{\sqrt{.155}} = .087$$



- Example #2: Keppel's Advertising data
  - What we found:
    - Data normally distributed
    - Homogeneity of variance
    - No outliers
  - What to do:
    - Conduct standard ANOVA

```
ONEWAY prefer BY group(1 3)
 /STAT=desc.
```

Descriptives

Preference for Ad

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Color Picture	7	4.7143	2.49762	.94401	2.4044	7.0242	1.00	8.00
Black & White Picture	7	6.1429	2.19306	.82890	4.1146	8.1711	3.00	9.00
No Picture	7	7.4286	1.71825	.64944	5.8395	9.0177	5.00	10.00
Total	21	6.0952	2.34318	.51132	5.0286	7.1618	1.00	10.00

$F(2,18) = 2.77, p = .09$

- Compute confidence intervals:

$$\bar{X}_{\cdot j} \pm \left(t_{crit}(df_W) * \sqrt{\frac{MSW}{n_j}} \right)$$

For 95% CI: $t_{crit}(18) = 2.10$

Color Picture: $4.71 \pm \left(2.10 * \sqrt{\frac{4.667}{7}} \right)$ (3.00, 6.42)

B&W Picture: $6.14 \pm \left(2.10 * \sqrt{\frac{4.667}{7}} \right)$ (4.43, 7.85)

No Picture: $7.42 \pm \left(2.10 * \sqrt{\frac{4.667}{7}} \right)$ (5.70, 9.43)

- Measures of effect size

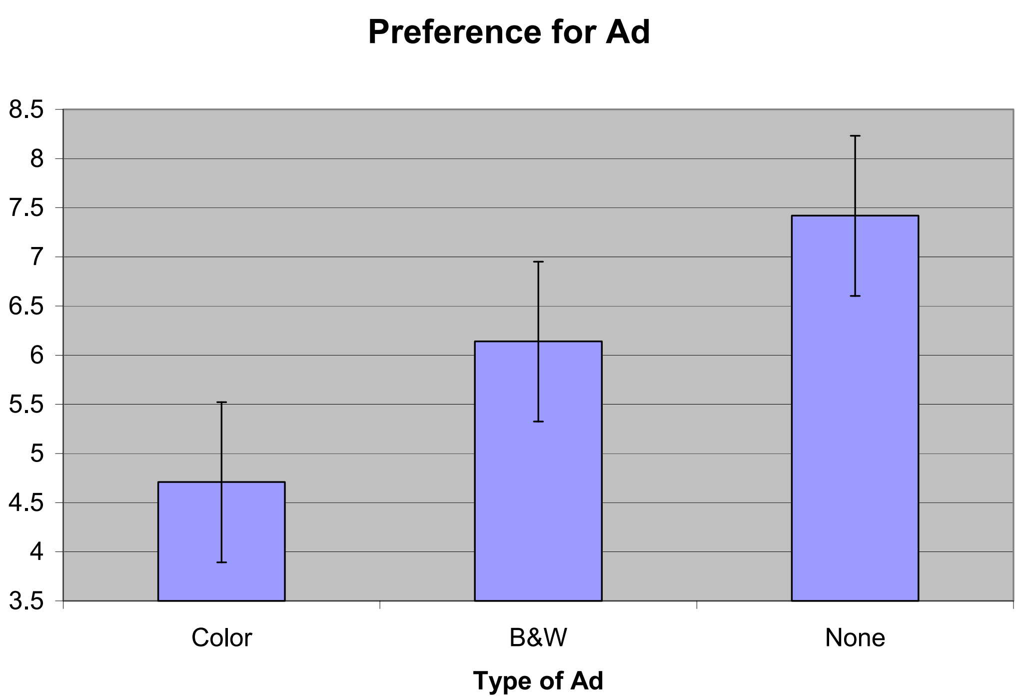
$$\hat{\omega}^2 = \frac{SSBetween - (a-1)MSWithin}{SSTotal + MSWithin}$$

$$\hat{\omega}^2 = \frac{25.81 - (2)4.667}{109.81 + 4.667} = .144$$

$$\hat{\sigma}_m = \sqrt{\frac{\sum_{j=1}^{a}(\mu._j - \mu..)^2}{a}}$$

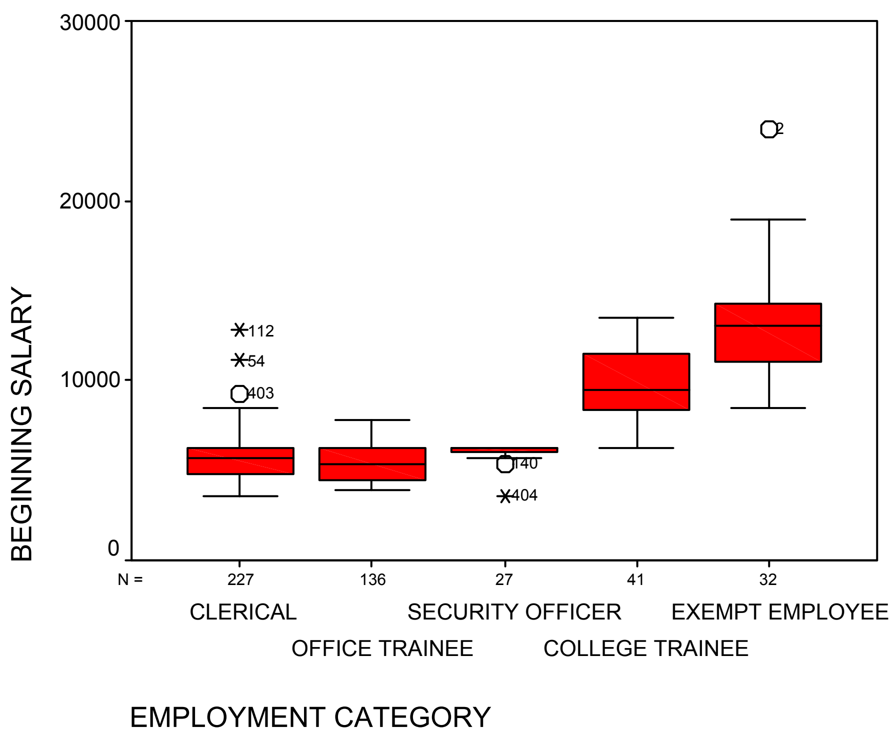
$$\hat{\sigma}_m = \sqrt{\frac{1.918 + .002 + 1.756}{3}} = 1.11$$

$$f = \frac{\hat{\sigma}_m}{\hat{\sigma}_e} = \frac{1.11}{\sqrt{4.667}} = .514$$



Note: Error Bars represent ± 1 Std Error

- Putting it all together: one more example.
- Example #4: Bank Data (from http://www.spss.com/tech/DataSets.html)
  - Data collected from 1969 to 1971 on 474 employees hired by a Midwestern bank.
  - Let's check to see if starting salary differs by position.



- Test Homogeneity of Variance:

Descriptives

	EMPLOYMENT		Statistic	Std. Error
BEGINNING SALARY	CLERICAL	Mean	5733.95	84.423
		Median	5700.00	
		Variance	1617876	
		Std. Deviation	1271.957	
		Interquartile Range	1500.00	
	OFFICE TRAINEE	Mean	5478.97	80.322
		Median	5400.00	
		Variance	877424.1	
		Std. Deviation	936.709	
		Interquartile Range	1800.00	
	SECURITY OFFICER	Mean	6031.11	103.248
		Median	6300.00	
		Variance	287825.6	
		Std. Deviation	536.494	
		Interquartile Range	300.00	
	COLLEGE TRAINEE	Mean	9956.49	311.859
		Median	9492.00	
		Variance	3987506	
		Std. Deviation	1996.874	
		Interquartile Range	3246.00	
	EXEMPT EMPLOYEE	Mean	13258.88	556.142
		Median	13098.00	
		Variance	9897415	
		Std. Deviation	3146.016	
		Interquartile Range	3384.00	

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
BEGINNING SALARY	Based on Mean	28.920	4	458	.000
	Based on Median	27.443	4	458	.000
	Based on Median and with adjusted df	27.443	4	175.027	.000
	Based on trimmed mean	28.390	4	458	.000

- Testing for normality in all five groups:

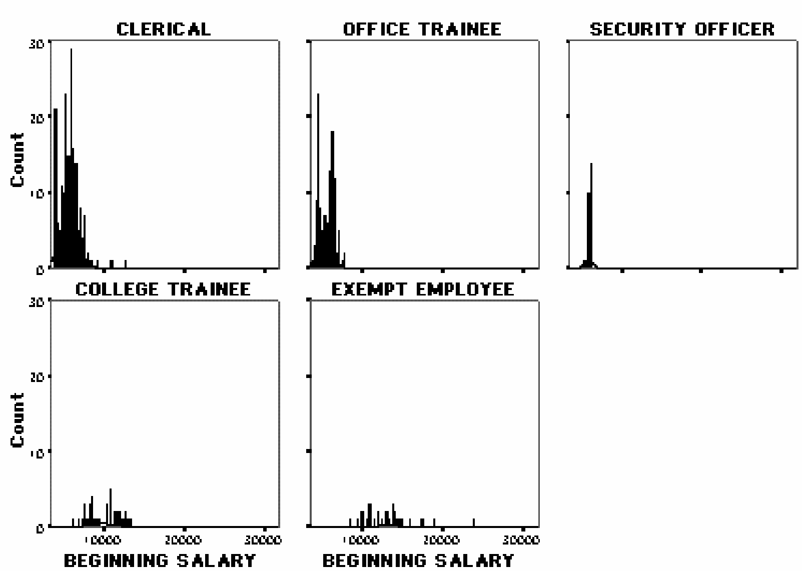
Descriptives

	EMPLOYMENT		Statistic	Std. Error
BEGINNING SALARY	CLERICAL	Mean	5733.95	84.423
		Median	5700.00	
		Skewness	1.251	.162
		Kurtosis	4.470	.322
	OFFICE TRAINEE	Mean	5478.97	80.322
		Median	5400.00	
		Skewness	.366	.208
		Kurtosis	-.939	.413
	SECURITY OFFICER	Mean	6031.11	103.248
		Median	6300.00	
		Skewness	-3.876	.448
		Kurtosis	17.203	.872
	COLLEGE TRAINEE	Mean	9956.49	311.859
		Median	9492.00	
		Skewness	.122	.369
		Kurtosis	-1.185	.724
	EXEMPT EMPLOYEE	Mean	13258.88	556.142
		Median	13098.00	
		Skewness	1.401	.414
		Kurtosis	3.232	.809

Tests of Normality

	EMPLOYMENT CATEGORY	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
BEGINNING SALARY	CLERICAL	.104	227	.000	.924	227	.000
	OFFICE TRAINEE	.148	136	.000	.924	136	.000
	SECURITY OFFICER	.366	27	.000	.499	27	.000
	COLLEGE TRAINEE	.158	41	.011	.947	41	.054
	EXEMPT EMPLOYEE	.155	32	.049	.903	32	.007

a. Lilliefors Significance Correction



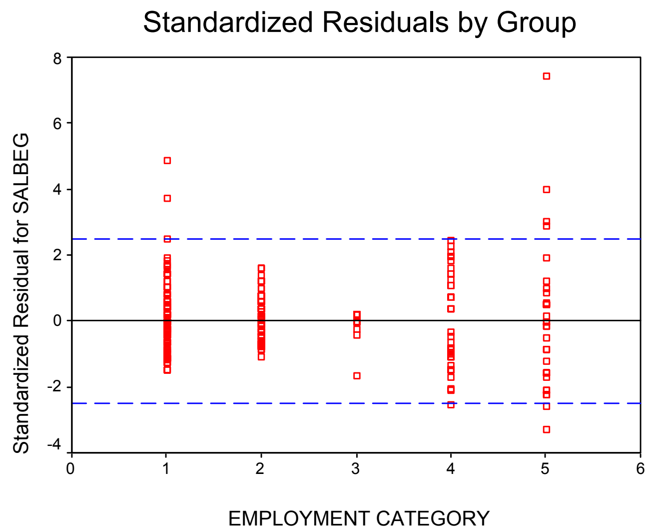
- Checking for outliers:

Extreme Values

	EMPLOYMENT CATEGOR`			Case Number	Value
Standardized Residual for SALBEG	CLERICAL	Highest	1	116	4.88
			2	58	3.71
			3	413	2.47
		Lowest	1	454	-1.48
			2	463	-1.48
			3	468	-1.48 [a]
	OFFICE TRAINEE	Highest	1		1.60
			2	263	1.60
			3	236	1.40 [b]
		Lowest	1	429	-1.09
			2	266	-.88
			3	214	-.88 [a]
	SECURITY OFFICER	Highest	1	421	.19
			2	405	.19
			3	117	.19 .[c]
		Lowest	1	414	-1.68
			2	146	-.44
			3	16	-.27 .[d]
	COLLEGE TRAINEE	Highest	1	17	2.45
			2	35	2.24
			3	6	2.10
		Lowest	1	306	-2.53
			2	334	-2.11
			3	305	-2.05 [a]
	EXEMPT EMPLOYEE	Highest	1	2	7.43
			2	67	3.97
			3	415	3.03
		Lowest	1	147	-3.29
			2	54	-2.60
			3	243	-2.26 .[e]

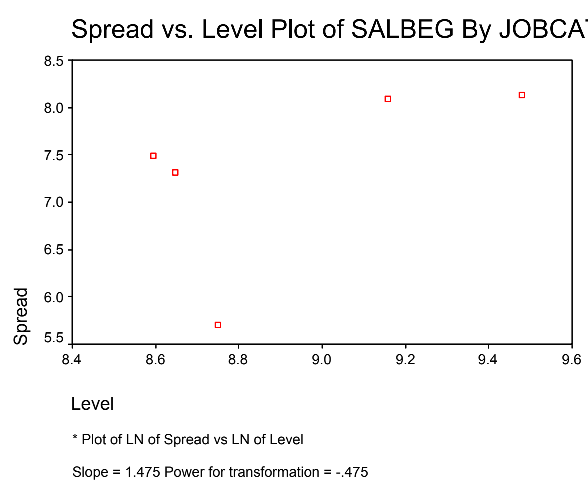
a.
b.
c.
d.
e.

- Eight-ish outliers??
  ($N = 463$)



- What we found:
  - Data non-normally distributed, possibly with different distributions in each group
  - Heterogeneity of Variances
  - 8-9 Outliers!
- What to do:
  - Transformation?
  - Brown-Forsythe Test (but this ignores the non-normality)

- Let's try a transformation first.
  - The spread and level plot may be helpful:



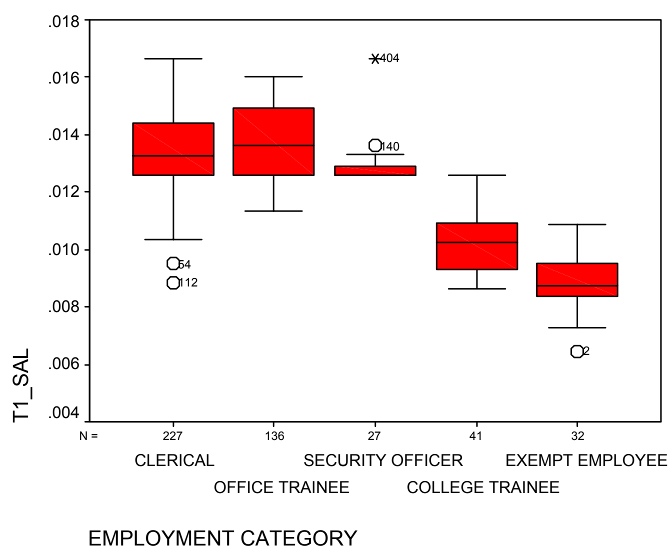
  - The plot recommends p = -.5 or inverse square root transformation

- We can also give Kirk's trick a shot:

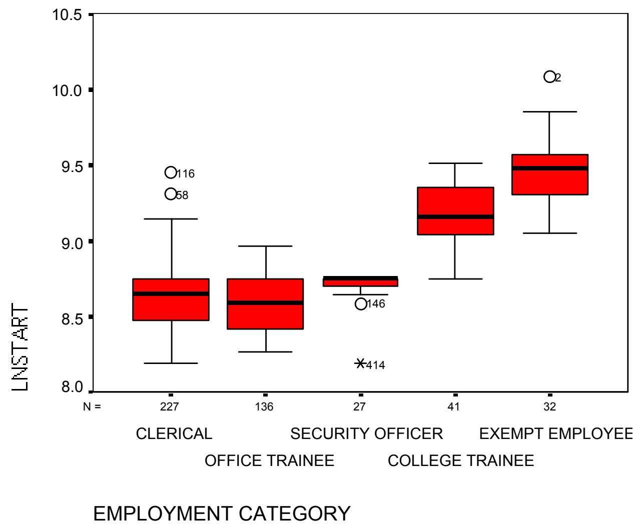
	Treatment Levels a_1	a_2	a_3	a_4	a_5	$\frac{Range_{largest}}{Range_{smallest}}$
Largest Score (L)	12792	7800	6300	13500	24000	
Smallest Score (S)	3600	3900	3600	6300	8496	
Range	9192	3900	2700	7200	15504	15504/2700 = 5.74
ln(L)	9.457	8.962	8.748	9.510	10.086	
ln(S)	8.189	8.269	8.189	8.748	9.047	
Range	1.268	0.693	0.560	0.762	1.038	1.268/.056 = 2.27
$\sqrt{L}$	113.102	88.318	79.373	116.190	154.919	
$\sqrt{S}$	60.000	62.450	60.000	79.373	92.174	
Range	53.102	25.868	19.373	36.817	62.746	53.10/19.37 = 2.74
1/L (* 10000)	0.782	1.282	1.587	0.741	0.417	
1/S (* 10000)	2.778	2.564	2.778	1.587	1.177	
Range (* 10000)	1.996	1.282	1.190	0.847	0.760	1.996/.760 = 2.63

- The three transformations are about equal, but the log transformation may be the best.

  - Let's go with the spread and level plot and try the inverse square root transformation.



- From the boxplot we can see that heterogeneity of variances is still a problem! Let's try the log transformation.



- Again, this transformation does not appear to solve the problem!

- We are left with the Brown-Forsythe Test (But to use this test, we must assume that the distributions at each level are relatively similar).

```
ONEWAY salbeg BY jobcat
  /STATISTICS DESCRIPTIVES BROWNFORSYTHE .
```

Descriptives

BEGINNING SALARY

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound
CLERICAL	227	5733.95	1271.957	84.423	5567.59	5900.30
OFFICE TRAINEE	136	5478.97	936.709	80.322	5320.12	5637.82
SECURITY OFFICER	27	6031.11	536.494	103.248	5818.88	6243.34
COLLEGE TRAINEE	41	9956.49	1996.874	311.859	9326.20	10586.78
EXEMPT EMPLOYEE	32	13258.88	3146.016	556.142	12124.62	14393.13
Total	463	6570.38	2626.953	122.085	6330.47	6810.29

ANOVA

BEGINNING SALARY

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2230311013.4	4	557577753.4	266.595	.000
Within Groups	957895695.66	458	2091475.318		
Total	3188206709.1	462			

Robust Tests of Equality of Means

BEGINNING SALARY

	Statistic[a]	df1	df2	Sig.
Brown-Forsythe	153.147	4	68.923	.000

a. Asymptotically F distributed.

- Report $F^*(4, 68.92) = 153.15$, $p < .001$

- Construct confidence intervals:
  - We found evidence for heterogeneity of variances, so we want to construct confidence intervals that take into account this heterogeneity.
  - In other words, the SPSS method of computing CIs is appropriate.

$$\bar{X}_{.j} \pm \left(t_{crit}(n_j - 1) * \frac{s_j}{\sqrt{n_j}} \right)$$

For j=1:

$$\bar{X}_{.j} = 5733.95$$

$$t_{crit}(226) = 1.9705$$

$$s_j = 1271.96$$

$$5733.95 \pm \left(1.9705 * \frac{1271.96}{\sqrt{227}} \right) \qquad (5567.59,\ 5900.31)$$

Descriptives

BEGINNING SALARY

	Mean	95% Confidence Interval for Mean	
		Lower Bound	Upper Bound
CLERICAL	5733.95	5567.59	5900.30
OFFICE TRAINEE	5478.97	5320.12	5637.82
SECURITY OFFICER	6031.11	5818.88	6243.34
COLLEGE TRAINEE	9956.49	9326.20	10586.78
EXEMPT EMPLOYEE	13258.88	12124.62	14393.13
Total	6570.38	6330.47	6810.29

- Compute effect sizes:

$$\hat{\omega}^2 = \frac{SSBetween - (a-1)MSWithin}{SSTotal + MSWithin}$$

$$\hat{\omega}^2 = \frac{2230311013 - (4)2091475}{3188206709 + 2091475} = .696$$

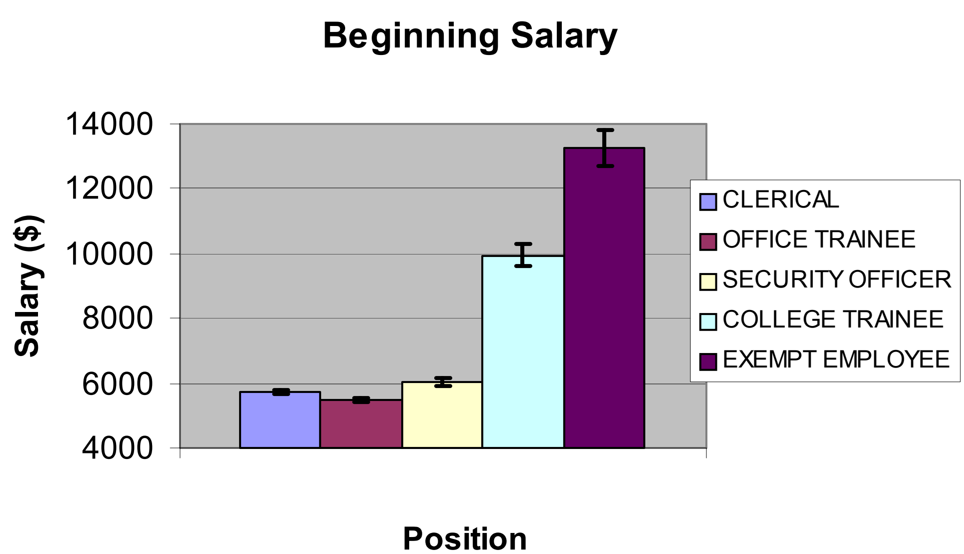
$$\hat{\sigma}_m = \sqrt{\frac{\sum_{j=1}^{a}(\mu_{.j} - \mu_{..})^2}{a}}$$

$$\hat{\sigma}_m = \sqrt{\frac{699620 + 1191175 + 290811 + 11465725 + 44735964}{5}} = 3417$$

$$f = \frac{\hat{\sigma}_m}{\hat{\sigma}_e} = \frac{3417}{\sqrt{2091475}} = 2.36$$

  - Note: When we compute the effect sizes, we make the homogeneity of variances assumption. It is not clear how valid these measures are (if at all) when we reject the homogeneity of variances assumption.

- Graph the data:
  - Because we rejected the homogeneity of variances assumption, use different error bars ($\pm$1 standard error) for each cell mean



Note: Error Bars represent $\pm$ 1 Std Error

