

# Model Assumptions

Andrew Zieffler

Educational Psychology

---

UNIVERSITY OF MINNESOTA

**Driven to Discover<sup>SM</sup>**

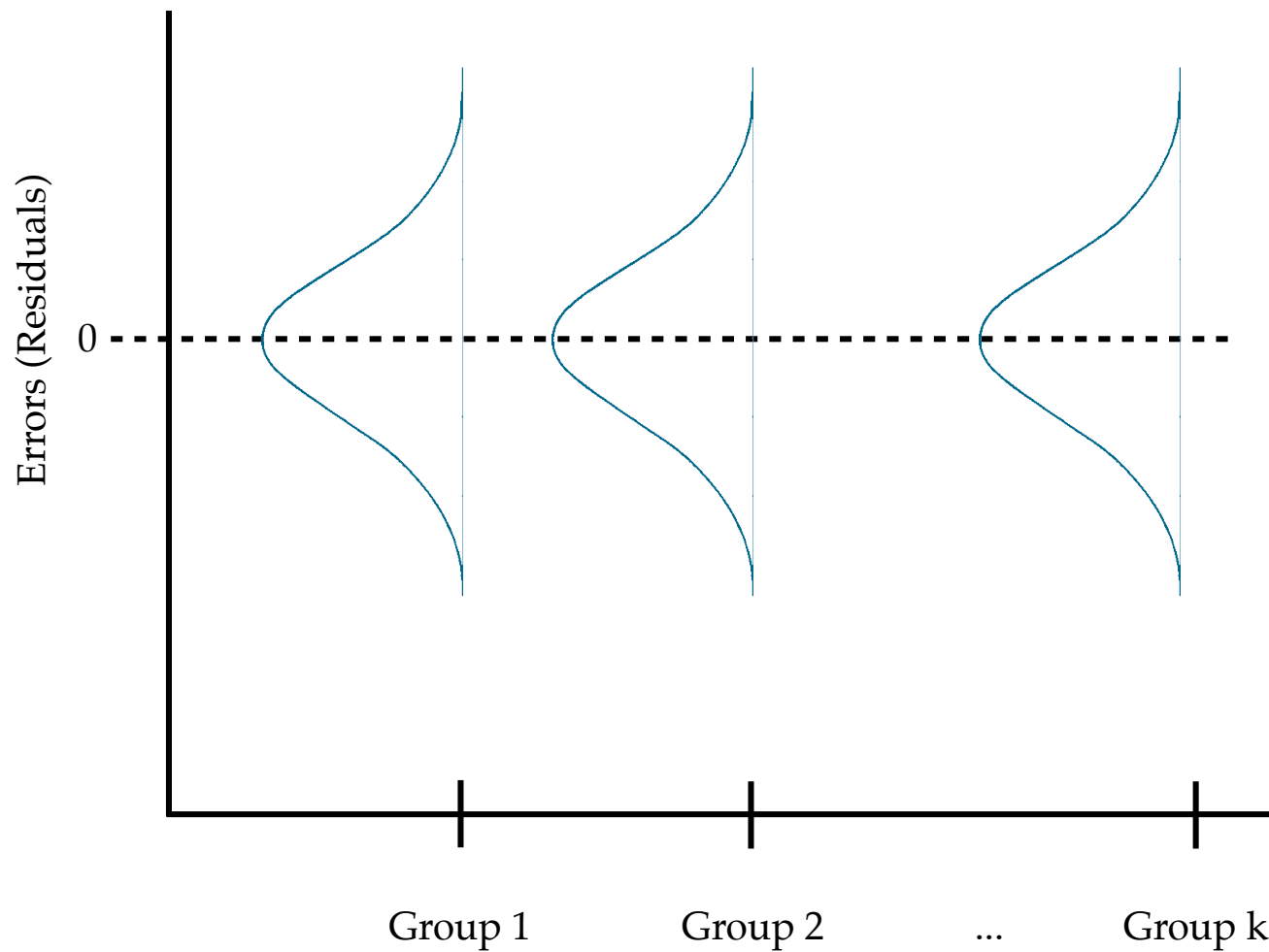
# Assumptions for the ANOVA Model

Both the marginal mean model and the conditional means model have assumptions about the errors if we are to believe the  $p$ -values.

**Violations of the assumptions affect the validity of the inferences.**

1. The errors (in the population) are **independent**.
2. The distribution of errors (in the population) is **normal** within each group.
3. All of the error distributions (in the population) have a **mean** of zero.
4. All of the error distributions (in the population) have the exact same variance (**homogeneity of variance**).

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$



Two important things to remember about model assumptions:

1. The assumptions are about the distribution of errors at each level of  $X$ .
2. The assumptions refer to the the distribution of errors in the population.

Recall that the sample  $e_i$  are approximations of the  $\varepsilon_i$

**Examining the  $e_i$  gives a good indication of how the  $\varepsilon_i$  behave...but**  
remember that sample data can deviate from what would be  
expected because they are a sample.

# Assumption: The Errors (in the Population) are Independent

The definition of independence relies on formal mathematics.

Loosely speaking **a set of observations is independent if knowing that one observation is above or below its mean value conveys no information about whether any other observation is above or below its mean value.** If observations are not independent, we say they are dependent or correlated.

Using a **random chance** in the study (to either select observations or assign them to levels of the predictor) will guarantee independence of the observations.

Assessing the independence assumption is primarily a logical argument.

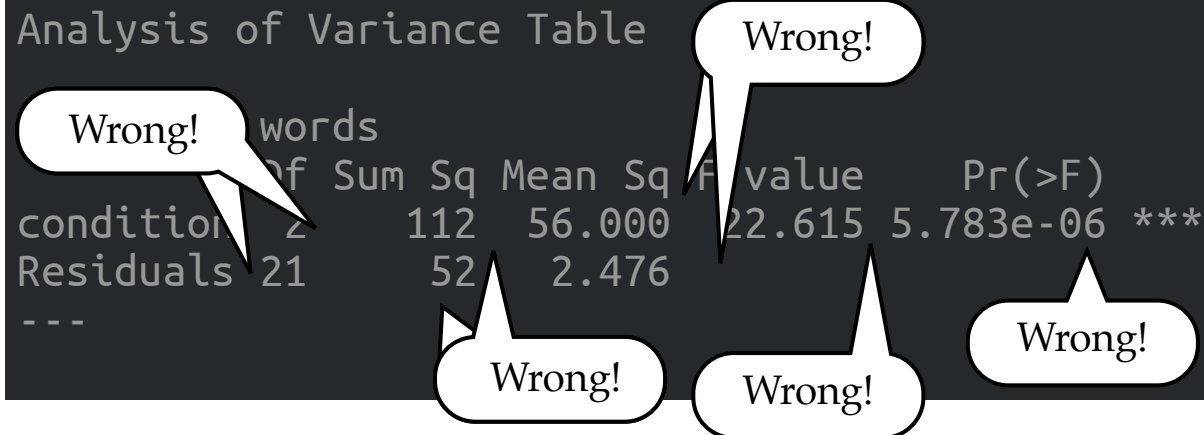
Aspects of data collection and analysis that **violate** independence:

- Physical (spatial) proximity in the collection of observations (e.g., convenience sampling based on location)
- Observations collected longitudinally (especially when they are the same subjects' data collected repeatedly)
- Analysis: When the level of assignment does not correspond to the level of analysis.

# What Happens if the Independence Assumption is Violated?

Violation of the independence assumption is a BIG problem.

Analysis of Variance Table



The table is an Analysis of Variance (ANOVA) table. It has five columns: 'Source', 'Df', 'Sum Sq', 'Mean Sq', and 'F value', and a final column for 'Pr(>F)'. The rows are 'words', 'condition', 'Residuals', and '---'. There are five callouts labeled 'Wrong!' pointing to: 1) the 'words' row, 2) the 'condition' row, 3) the 'Residuals' row, 4) the 'F value' column, and 5) the 'Pr(>F)' column.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
words					
condition	2	112	56.000	22.615	5.783e-06 ***
Residuals	21	52	2.476		
---					

**What to do:** Use a method for correlated (non-independent) data  
(Take EPsy 8252 to find out more!)

Assumption:  $\varepsilon_{ij} \sim N(0, \sigma^2)$

These assumptions are about the distribution of errors for each group. This assumption has three parts to examine:

- Each distribution is normal
- Each distribution has a mean of 0
- Each distribution has the same variance

We will use the `fortify()` function from the **ggplot2** library to obtain the sample errors and evaluate the assumption.

```
# fortify the model  
> outa = fortify(lm.a)  
> head(outa)
```

These are the errors.

	words	condition	...	.fitted	.resid	.stdresid
1	7	30s	...	6	1.000000e+00	0.6793662
2	3	30s	...	6	-3.000000e+00	-2.0380987
3	6	30s	...	6	-9.15934e-16	0.0000000
4	6	30s	...	6	0.000000e+00	0.0000000
5	5	30s	...	6	-1.000000e+00	-0.6793662
6	8	30s	...	6	2.000000e+00	1.3587324

To evaluate whether the mean of each error distribution is zero, we can compute summary measures using the `describeBy()` function.

```
> describeBy(outa$resid, outa$condition)
```

```
group: Private
```

```
group: 30s
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
1	1	8	0	1.51	0	0	1.48	-3	2	5	-0.65	-0.61	0.53

-----

```
group: 60s
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
1	1	8	0	1.41	0.5	0	0.74	-3	1	4	-1.06	-0.31	0.5

-----

```
group: 180s
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
1	1	8	0	1.77	0	0	1.48	-3	3	6	0	-0.9	0.63

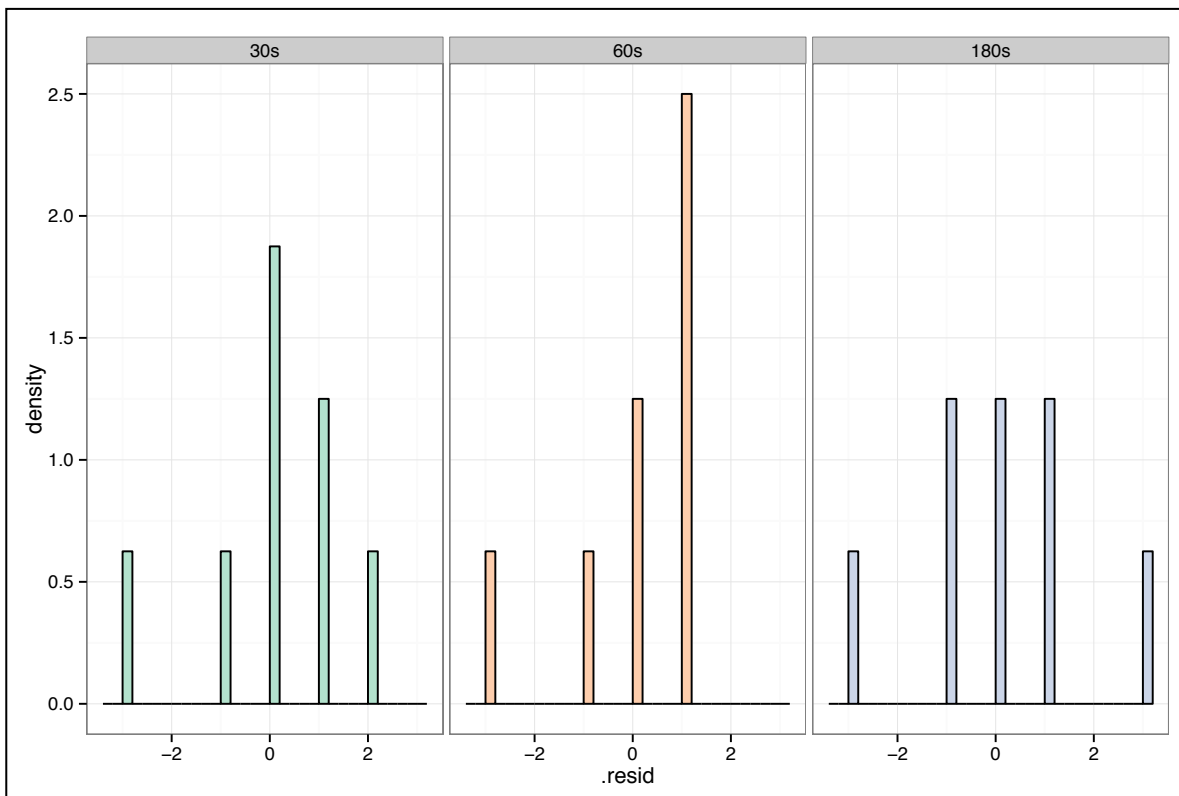
The mean of the errors/  
residuals for each group is 0.

Because we use least squares estimation to compute the estimates, the sample errors will always have a mean of zero for each distribution.



To evaluate the normality assumption, we examine plots of the sample error distributions. Histograms or density plots are useful plots to examine the shape of a distribution.

```
# Plot the error distribution for each group
> ggplot(data = outa, aes(x = .resid, y = ..density.., fill = condition)) +
  geom_histogram(color = "black") +
  facet_wrap(~ condition) +
  scale_fill_brewer(palette = "Pastel2", guide = FALSE) +
  theme_bw()
```



The errors for the 30s and 180s groups seem relatively normally distributed, but those for the 60s group are negatively skewed.

Is the error distribution in the population skewed?

Or, is the error distribution in the population normally distributed?

To evaluate the homogeneity of variance assumption, we typically compare the variances of the sample error distributions. To get the variances, we can square the SDs from the describeBy() output.

```
# Compute the variance for the 30s group
> 1.51 ^ 2
[1] 2.28

# Compute the variance for the 60s group
> 1.41 ^ 2
[1] 1.99

# Compute the variance for the 180s group
> 1.77 ^ 2
[1] 3.13
```

The sample variances are not equal, but that may be because of sampling error...remember, the assumptions are about the population errors!

Some people also use a test of the variances called Levene's test.

$$H_0 : \sigma_{\epsilon_1}^2 = \sigma_{\epsilon_2}^2 = \dots = \sigma_{\epsilon_k}^2$$

```
# Load car library (may need to install it first)
> library(car)

> leveneTest(lm.a)

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.1373 0.8725
      21
```

This test tends to reject more often than it should when the null hypothesis is true (makes too many type I errors)

Better to make this evaluation from plots.

# Have the Model Assumptions been Met?

## Independence



Mean of each error distribution is zero



30s



60s



180s

Each error distribution is normally distributed



30s



60s



180s

Homogeneity of variance



In practice, it turns out that the tests used for ANOVA (and the  $t$ -test) are still okay to use if one of these last two assumptions has been violated.

# How to Proceed if the Normality or Homogeneity of Variance Assumption have been Violated

Are the group's sample sizes balanced?



**Yes**

- Violations of the normality assumption are generally not problematic
- Minor violations of the homogeneity assumption are generally not problematic
  - ▶ If the largest and smallest variances are in a ratio of 4:1 (or less), ANOVA generally is robust for non-small sample sizes

**Take Home Message:** With equal group sample sizes ANOVA is generally not problematic, even if normality or homogeneity of variance have been violated.

## No

- Violations of the normality assumption are sometimes problematic
  - ▶ If the distributions are non-normal, but have the same shape, ANOVA is relatively robust
  - ▶ Differences in kurtosis are generally less problematic than differences in skewness
- Violations in homogeneity of variance can be problematic at ratios of 2:1 or larger

**Take Home Message:** For unbalanced designs, ANOVA can be very problematic.

Small sample sizes make things worse.

## Potential Solutions

- Try a transformation of the data
- Use a nonparametric method to analyze the data
- Use a simulation method (e.g., bootstrapping)
- Use a method of analysis for unequal variances

Johnson, C. C. (1993). The effects of violation of data set assumptions when using the oneway, fixed effects analysis of variance and the one concomitant analysis of covariance statistical procedures. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans, LA. <http://eric.ed.gov/?id=ED365720>

Even though the sample error distribution for the 60s group was not normally distributed, we will say the model's assumptions have been reasonably satisfied.

Since we have balanced sample sizes, even if the normality assumption were violated, we can probably use the ANOVA and believe the  $p$ -values.