

# Lesson 5:

# Analysis of Variance

# **This Lesson's Goals**

Learn about analysis of variance (ANOVA)

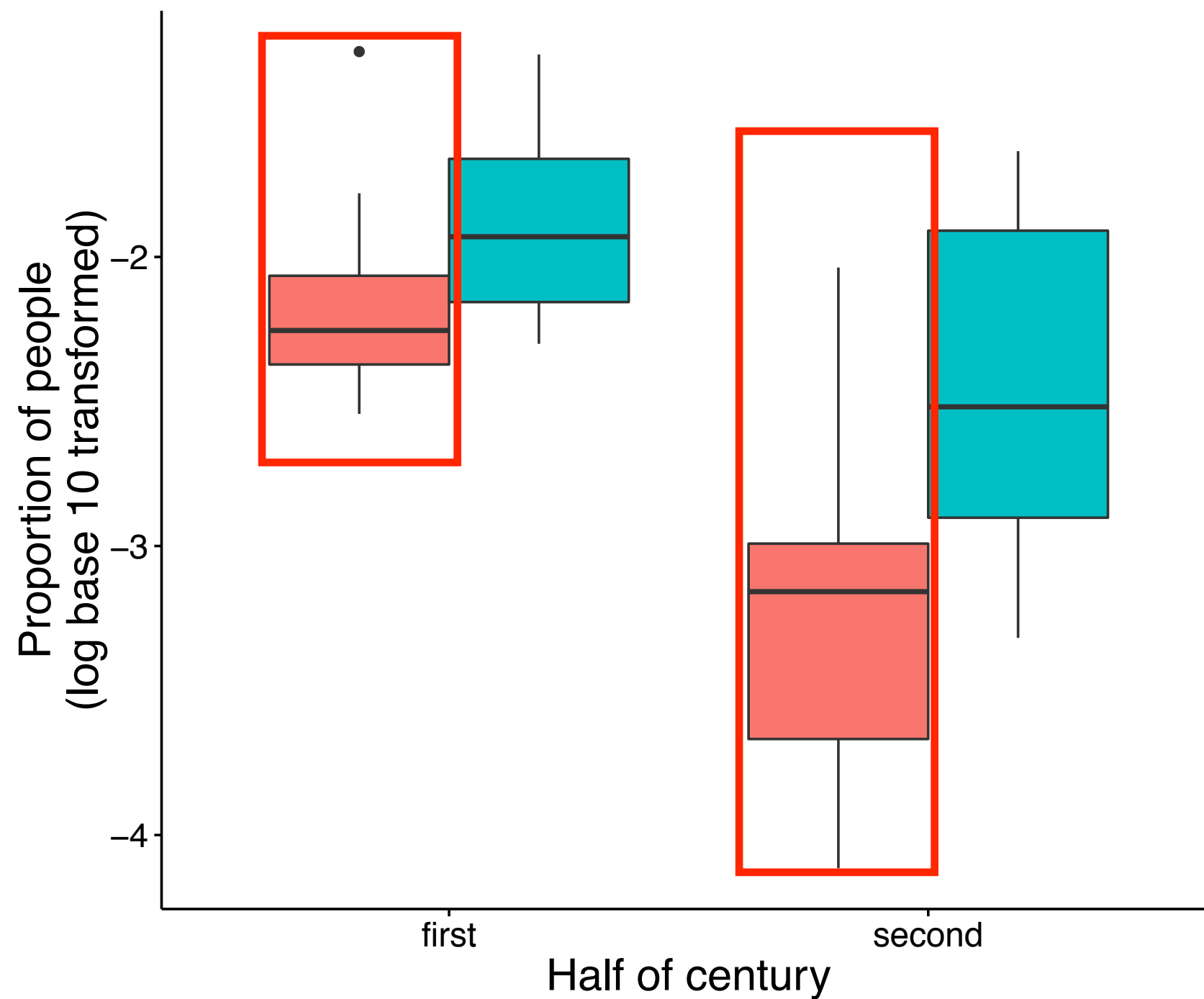
Make figures for data for an ANOVA

Do an ANOVA in R

Summarise results in an R Markdown document

# Proportion of People with Popular Names for 1901

sex female male



But I don't want to know the effect of century half ***only*** for females?

I want it for the whole data set!

**Let's run an ANOVA.**

# Math (Part 1)

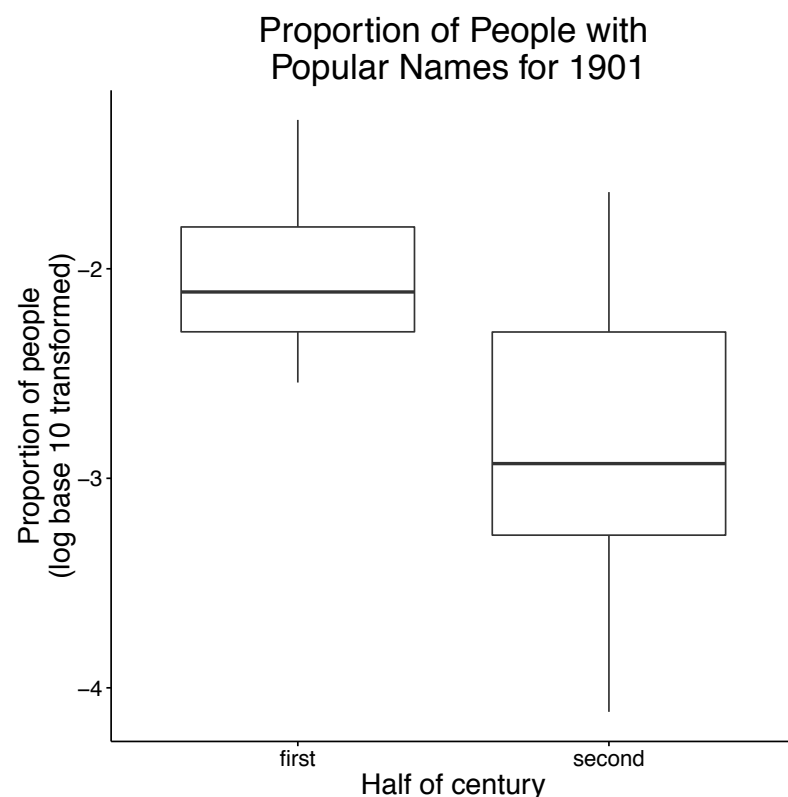
## Details

This provides a wrapper to `lm` for fitting linear models to balanced or unbalanced experimental designs.

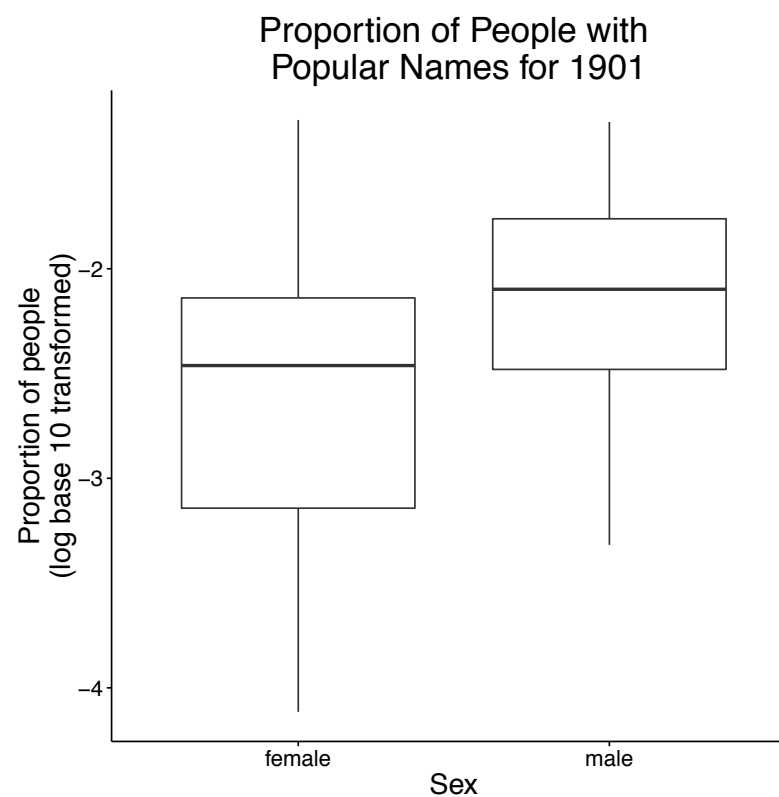
The main difference from `lm` is in the way `print`, `summary` and so on handle the fit: this is expressed in the traditional language of the analysis of variance rather than that of linear models.

*R help*

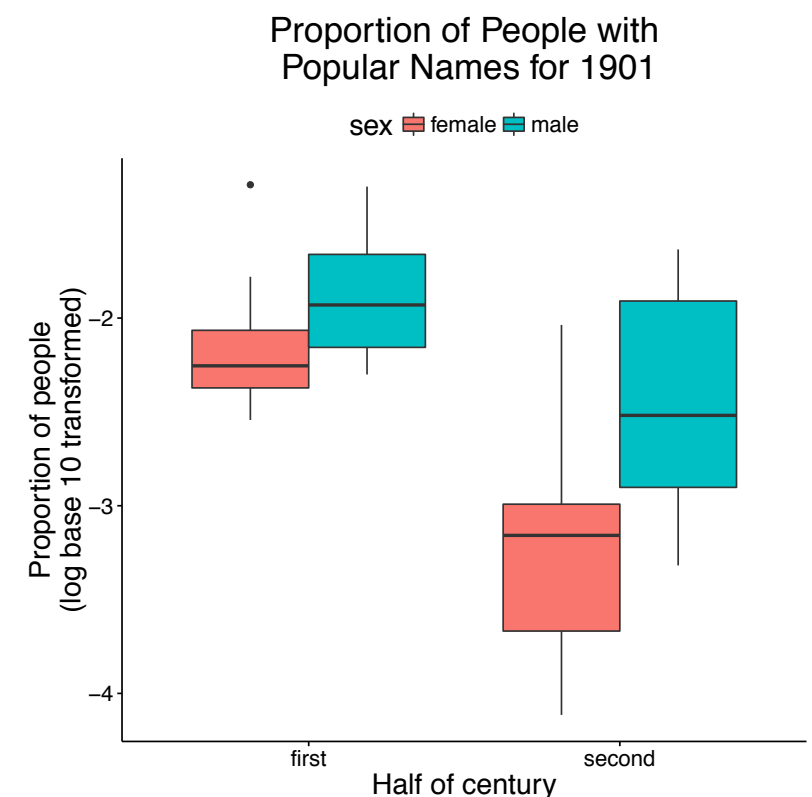
### main effect of $x_1$



### main effect of $x_2$



### interaction $x_1 \times x_2$



# partition of sum of squares

Mathematically, the sum of squared deviations is an unscaled, or unadjusted **measure of dispersion (also called variability)**. When scaled for the number of degrees of freedom, it estimates the variance, or spread of the observations about their mean value. Partitioning of the sum of squared deviations into various components allows the overall variability in a dataset to be ascribed to **different types or sources of variability**, with the relative importance of each being quantified by the size of each component of the overall sum of squares.

*Wikipedia: Partition of sums of squares*

$y$  = continuous dependent variable  
 $X_n$  = independent variable(s)

error variable = e.g. subject, item  
averaged variable = e.g. subject, item

error  
variable

$X_1$

$X_2$

$y$

averaged  
over *years*



<b>name</b>	<b>century_half</b>	<b>sex</b>	<b>prop</b>
Albert	first	male	-2.150820
Albert	second	male	-2.897882
Alice	first	female	-2.097719
Alice	second	female	-3.123840
...	...	...	....
Willie	first	male	-2.175477
Willie	second	male	-2.912281



$y$  = continuous dependent variable  
 $X_n$  = independent variable(s)

error variable = e.g. subject, item  
averaged variable = e.g. subject, item

error  
variable

$X_1$

$X_2$

averaged  
over *names*



$y$

<b>year</b>	<b>century_half</b>	<b>sex</b>	<b>prop</b>
1901	first	female	-1.904843
1901	first	male	-1.751544
1902	first	female	-1.903983
1902	first	male	-1.755954
...	...	...	....
2000	second	female	-3.360910
2000	second	male	-2.891397

# R Code (Part 1)

$$y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + e_i$$

`aov(prop_log10_mean ~ century_half * sex)`  
`anova(popnames_interaction.lm)`

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
century_half	1	12.997	12.997	60.309	3.09e-11	***
sex	1	5.256	5.256	24.389	4.55e-06	***
century_half:sex	1	0.932	0.932	4.325	0.0409	*
Residuals	76	16.378	0.215			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## linear model without interaction

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-2.27486	0.09182	-24.776	< 2e-16 ***
century_halfsecond	-0.80612	0.10602	-7.603	5.88e-11 ***
sexmale	0.51263	0.10602	4.835	6.66e-06 ***

## linear model with interaction

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-2.1669	0.1038	-20.875	< 2e-16 ***
century_halfsecond	-1.0220	0.1468	-6.962	1.04e-09 ***
sexmale	0.2967	0.1468	2.021	0.0468 *
century_halfsecond:sexmale	0.4318	0.2076	2.080	0.0409 *

## analysis of variance (ANOVA)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
century_half	1	12.997	12.997	60.309	3.09e-11 ***
sex	1	5.256	5.256	24.389	4.55e-06 ***
century_half:sex	1	0.932	0.932	4.325	0.0409 *
Residuals	76	16.378	0.215		

## linear model with interaction

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-2.1669	0.1038	-20.875	< 2e-16	***
century_halfsecond	-1.0220	0.1468	-6.962	1.04e-09	***
sexmale	0.2967	0.1468	2.021	0.0468	*
century_halfsecond:sexmale	0.4318	0.2076	2.080	0.0409	*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4642 on 76 degrees of freedom

Multiple R-squared: 0.5395, Adjusted R-squared: 0.5213

F-statistic: 29.67 on 3 and 76 DF, p-value: 8.345e-13

## analysis of variance (ANOVA)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
century_half	1	12.997	12.997	60.309	3.09e-11	***
sex	1	5.256	5.256	24.389	4.55e-06	***
century_half:sex	1	0.932	0.932	4.325	0.0409	*
Residuals	76	16.378	0.215			

mean = 29.67433



But, what about within- and between-factor ANOVAs?

How to I do that?

**With an 'Error' term.**

# Math (Part 2)



## Details

This provides a wrapper to `lm` for fitting linear models to balanced or unbalanced experimental designs.

The main difference from `lm` is in the way `print`, `summary` and so on handle the fit: this is expressed in the traditional language of the analysis of variance rather than that of linear models.



*R help*



by-name

name	century_half	sex	prop
Albert	first	male	-2.150820
Albert	second	male	-2.897882
Alice	first	female	-2.097719
Alice	second	female	-3.123840
...	...	...	....
Willie	first	male	-2.175477
Willie	second	male	-2.912281

by-year

year	century_half	sex	prop
1901	first	female	-1.904843
1901	first	male	-1.751544
1902	first	female	-1.903983
1902	first	male	-1.755954
...	...	...	....
2000	second	female	-3.360910
2000	second	male	-2.891397

# **R Code (Part 2)**

```
aov(prop_log10_mean ~ century_half * sex
    + Error(name/century_half))
```

Error: name

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	5.256	5.256	14.21	0.000556 ***
Residuals	38	14.056	0.370		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Error: name:century\_half

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
century_half	1	12.997	12.997	212.71	< 2e-16 ***
century_half:sex	1	0.932	0.932	15.26	0.000373 ***
Residuals	38	2.322	0.061		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## ANOVA without error term

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
century_half	1	12.997	12.997	60.309	3.09e-11	***
sex	1	5.256	5.256	24.389	4.55e-06	***
century_half:sex	1	0.932	0.932	4.325	0.0409	*
Residuals	76	16.378	0.215			

## ANOVA with error term

Error: name

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
sex	1	5.256	5.256	14.21	0.000556	***
Residuals	38	14.056	0.370			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Error: name:century\_half

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
century_half	1	12.997	12.997	212.71	< 2e-16	***
century_half:sex	1	0.932	0.932	15.26	0.000373	***
Residuals	38	2.322	0.061			

But, I have different numbers in my two groups and SPSS gives me different values than R.

What do I do to get the SPSS values?

**Run an ANOVA with ezANOVA.**

# Math (Part 3)

$$y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + e_i$$

**Type I:** Order of variables matters. Sum of Squares for  $x_1$  is computed, but not controlling for  $x_2$ , which may be a problem if  $x_2$  is unbalanced.

**Type II:** Only to be used if there is no interaction in the model.

**Type III:** Order of variables *does not* matter. Sum of Squares for  $x_1$  and  $x_2$  is computed as if both were included last, thus accounting for if the other variable is unbalanced.

$SS(x_1 | x_2)$   
 sum of squares of  $x_1$  given  $x_2$

	$x_1$	$x_2$	$x_1 x_2$
Type 1	$SS(x_1)$	$SS(x_2 x_1)$	$SS(x_1 x_2 x_1, x_2)$
Type 2	$SS(x_1 x_2)$	$SS(x_2 x_1)$	$SS(x_1 x_2 x_1, x_2)$
Type 3	$SS(x_1 x_2, x_1 x_2)$	$SS(x_2 x_1, x_1 x_2)$	$SS(x_1 x_2 x_1, x_2)$

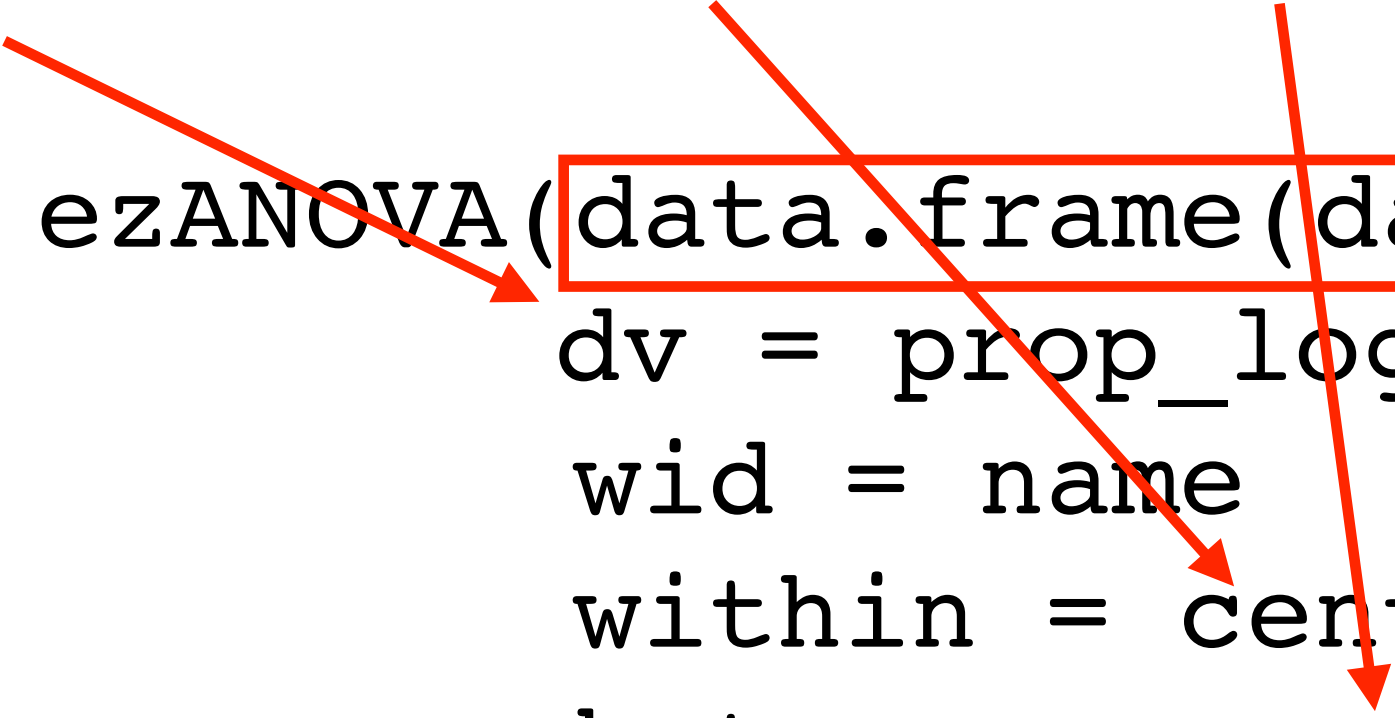


**R Code (Part 3)**

ez

$$y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + e_i$$

```
ezANOVA(data.frame(data_names),  
         dv = prop_log10_mean,  
         wid = name  
         within = century_half,  
         between = sex,  
         type = 3)
```



```
ez      ezANOVA(data.frame(data_names),  
                dv = prop_log10_mean,  
                wid = name  
                within = century_half,  
                between = sex,  
                type = 3)
```

`dv` = dependent variable

`wid` = error term

`within` = any variable(s) that is(are) within the wid

`between` = any variable(s) that is(are) between the wid

`type` = Sum of Squares type (1, 2, or 3)

type

Numeric value (either 1, 2 or 3) specifying the Sums of Squares “type” to employ when data are unbalanced (eg. when group sizes differ). type = 2 is the default because this will yield identical ANOVA results as type = 1 when data are balanced but type = 2 will additionally yield various assumption tests where appropriate. When data are unbalanced, users are warned that they should give special consideration to the value of type. type=3 will emulate the approach taken by popular commercial statistics packages like SAS and SPSS but users are warned that this approach is not without criticism.

*ez help*

**ez**

$$y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + e_i$$

```
ezANOVA(data.frame(data_names),  
         dv = prop_log10_mean,  
         wid = name  
         within = century_half,  
         between = sex,  
         type = 3)
```

\$ANOVA

	Effect	DFn	DFd	F	p	p<.05	ges
1	sex	1	38	14.20865	5.560671e-04	*	0.24294331
2	century_half	1	38	212.71271	3.754057e-17	*	0.44244482
3	sex:century_half	1	38	15.25534	3.730506e-04	*	0.05384696

# ANOVA with error term

Error: name								Error: name:century_half							
	Df	Sum Sq	Mean Sq	F value	Pr(>F)				Df	Sum Sq	Mean Sq	F value	Pr(>F)		
sex	1	5.256	5.256	14.21	0.000556 ***	century_half			1	12.997	12.997	212.71	< 2e-16 ***		
Residuals	38	14.056	0.370			century_half:sex			1	0.932	0.932	15.26	0.000373 ***		
						Residuals			38	2.322	0.061				

## ezANOVA with type 1 Sum of Squares

\$ANOVA									
	Effect	DFn	DFd	F	p	p<.05		ges	
1	sex	1	38	14.20865	5.560671e-04	*	0.24294331		
2	century_half	1	38	212.71271	3.754057e-17	*	0.44244482		
3	sex:century_half	1	38	15.25534	3.730506e-04	*	0.05384696		

## ezANOVA with type 2 Sum of Squares

\$ANOVA									
	Effect	DFn	DFd	F	p	p<.05		ges	
2	sex	1	38	14.20865	5.560671e-04	*	0.24294331		
3	century_half	1	38	212.71271	3.754057e-17	*	0.44244482		
4	sex:century_half	1	38	15.25534	3.730506e-04	*	0.05384696		

## ezANOVA with type 3 Sum of Squares

\$ANOVA									
	Effect	DFn	DFd	F	p	p<.05		ges	
2	sex	1	38	14.20865	5.560671e-04	*	0.24294331		
3	century_half	1	38	212.71271	3.754057e-17	*	0.44244482		
4	sex:century_half	1	38	15.25534	3.730506e-04	*	0.05384696		

Same analysis, top **20** names for females but top **18** for males.



# ANOVA with error term

Error: name								Error: name:century_half					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)			Df	Sum Sq	Mean Sq	F value	Pr(>F)	
sex	1	5.353	5.353	14.14	0.000603 ***	century_half		1	12.941	12.941	207.83	< 2e-16 ***	
Residuals	36	13.629	0.379			century_half:sex		1	0.817	0.817	13.12	0.000893 ***	
						Residuals		36	2.242	0.062			

## ezANOVA with type 1 Sum of Squares

\$ANOVA									
	Effect	DFn	DFd	F	p	p<.05		ges	
1	sex	1	36	14.13910	6.031431e-04	*	0.25221467		
2	century_half	1	36	207.83060	1.582563e-16	*	0.44914616		
3	sex:century_half	1	36	13.12093	8.929599e-04	*	0.04895614		

## ezANOVA with type 2 Sum of Squares

\$ANOVA									
	Effect	DFn	DFd	F	p	p<.05		ges	
2	sex	1	36	14.13910	6.031431e-04	*	0.25221467		
3	century_half	1	36	207.83060	1.582563e-16	*	0.44914616		
4	sex:century_half	1	36	13.12093	8.929599e-04	*	0.04895614		

## ezANOVA with type 3 Sum of Squares

\$ANOVA									
	Effect	DFn	DFd	F	p	p<.05		ges	
2	sex	1	36	14.13910	6.031431e-04	*	0.25221467		
3	century_half	1	36	201.80201	2.488668e-16	*	0.44187463		
4	sex:century_half	1	36	13.12093	8.929599e-04	*	0.04895614		



Same analysis, sexes balanced, but missing **4** (2 F, 2 M) data points in second half of century.

**NOTE: Can't have any variables as "within" now since that's not true for all names.**

# ANOVA with no term

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
century_half	1	11.789	11.789	54.167	2.39e-10	***
sex	1	4.655	4.655	21.387	1.61e-05	***
century_half:sex	1	0.830	0.830	3.811	0.0548	.
Residuals	72	15.671	0.218			

## ezANOVA with type 1 Sum of Squares

\$ANOVA

	Effect	DFn	DFd	F	p	p<.05	ges
1	century_half	1	72	54.167311	2.386530e-10	*	0.42932920
2	sex	1	72	21.387471	1.612390e-05	*	0.22901863
3	century_half:sex	1	72	3.811294	5.479694e-02		0.05027344

## ezANOVA with type 2 Sum of Squares

\$ANOVA

	Effect	DFn	DFd	F	p	p<.05	ges
1	century_half	1	72	54.167311	2.386530e-10	*	0.42932920
2	sex	1	72	21.387471	1.612390e-05	*	0.22901863
3	century_half:sex	1	72	3.811294	5.479694e-02		0.05027344

\$`Levene's Test for Homogeneity of Variance`

	DFn	DFd	SSn	SSd	F	p	p<.05
1	3	72	1.07691	5.012476	5.156304	0.002762266	*

## ezANOVA with type 3 Sum of Squares

\$ANOVA

	Effect	DFn	DFd	F	p	p<.05	ges
2	century_half	1	72	54.167311	2.386530e-10	*	0.42932920
3	sex	1	72	22.287835	1.125253e-05	*	0.23638081
4	century_half:sex	1	72	3.811294	5.479694e-02		0.05027344

\$`Levene's Test for Homogeneity of Variance`

	DFn	DFd	SSn	SSd	F	p	p<.05
1	3	72	1.07691	5.012476	5.156304	0.002762266	*

# **When Can I (for sure) Run an ANOVA?**

Continuous dependent variable

Balanced data set (same number of data points in all cells)

Same variance within each group analyzed

Samples are independently drawn

## **When Can I NOT Run an ANOVA?**

Count data for dependent variable

Are you taking the percentage of count data (i.e. percent correct)? That's still count data.

**Lab**





*Battle Star Galactica*



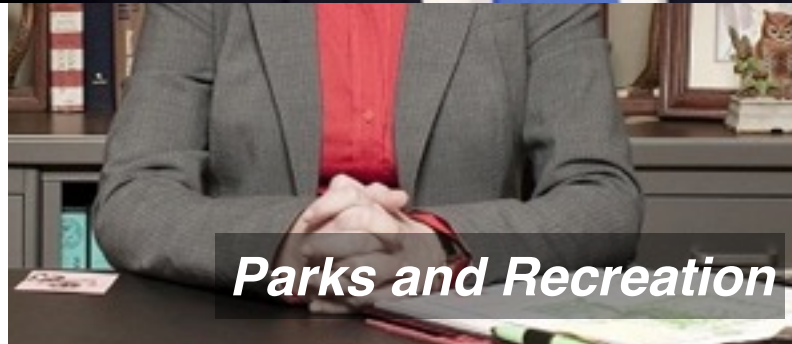
*Veep*



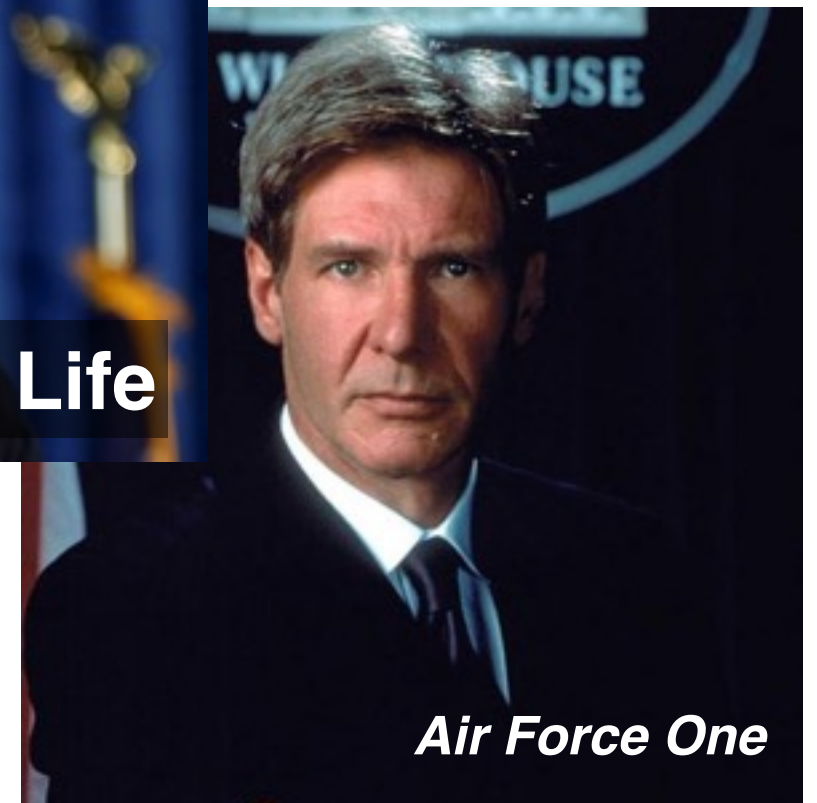
**Real Life**



*Independence Day*



*Parks and Recreation*



*Air Force One*

## **Dataset:** United States Presidential Elections

Incumbent Party: Do Democrats or Republicans get a higher percentage of the vote when they are an incumbent?

*source: The American Presidency Project*



1861

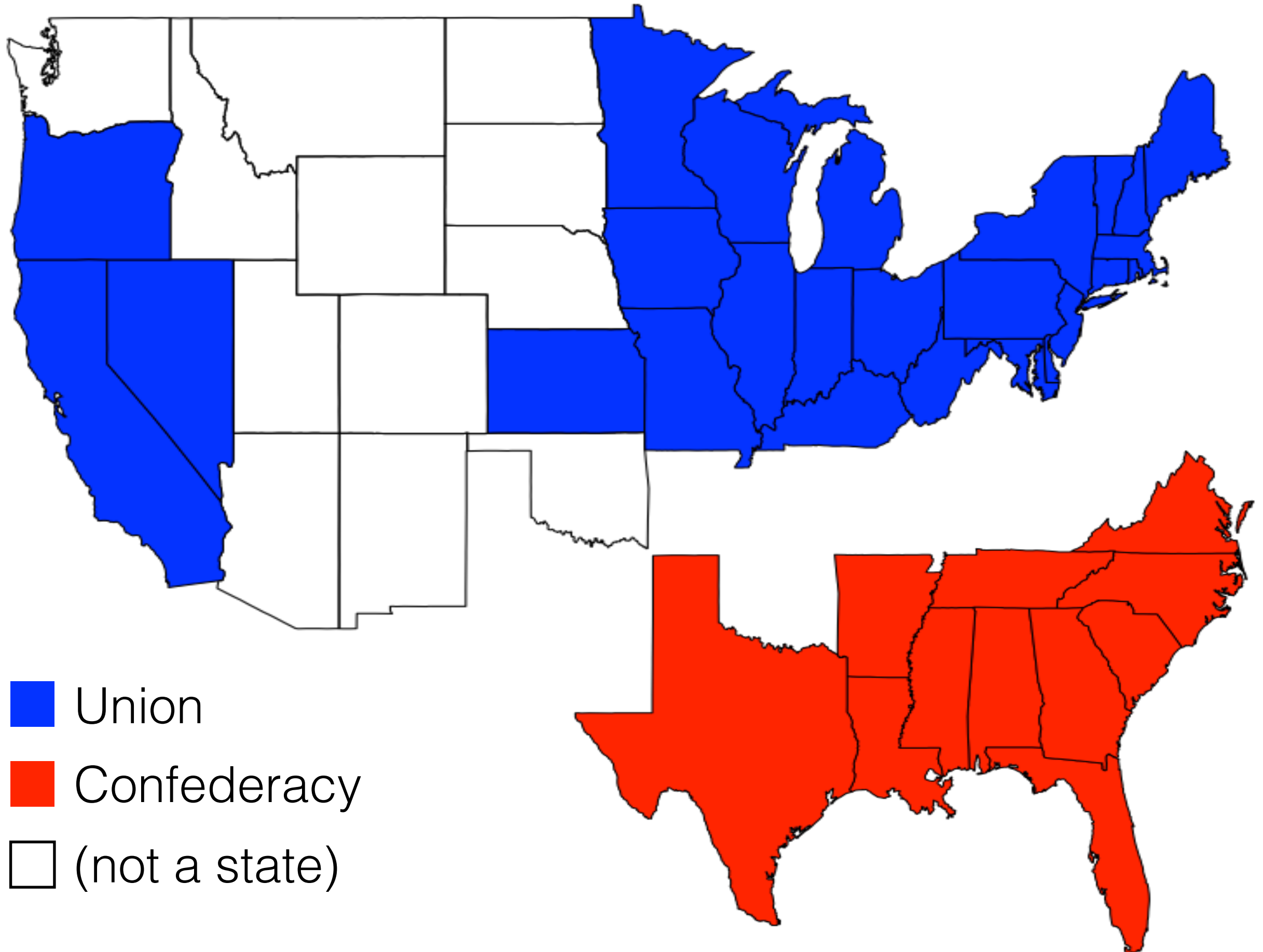




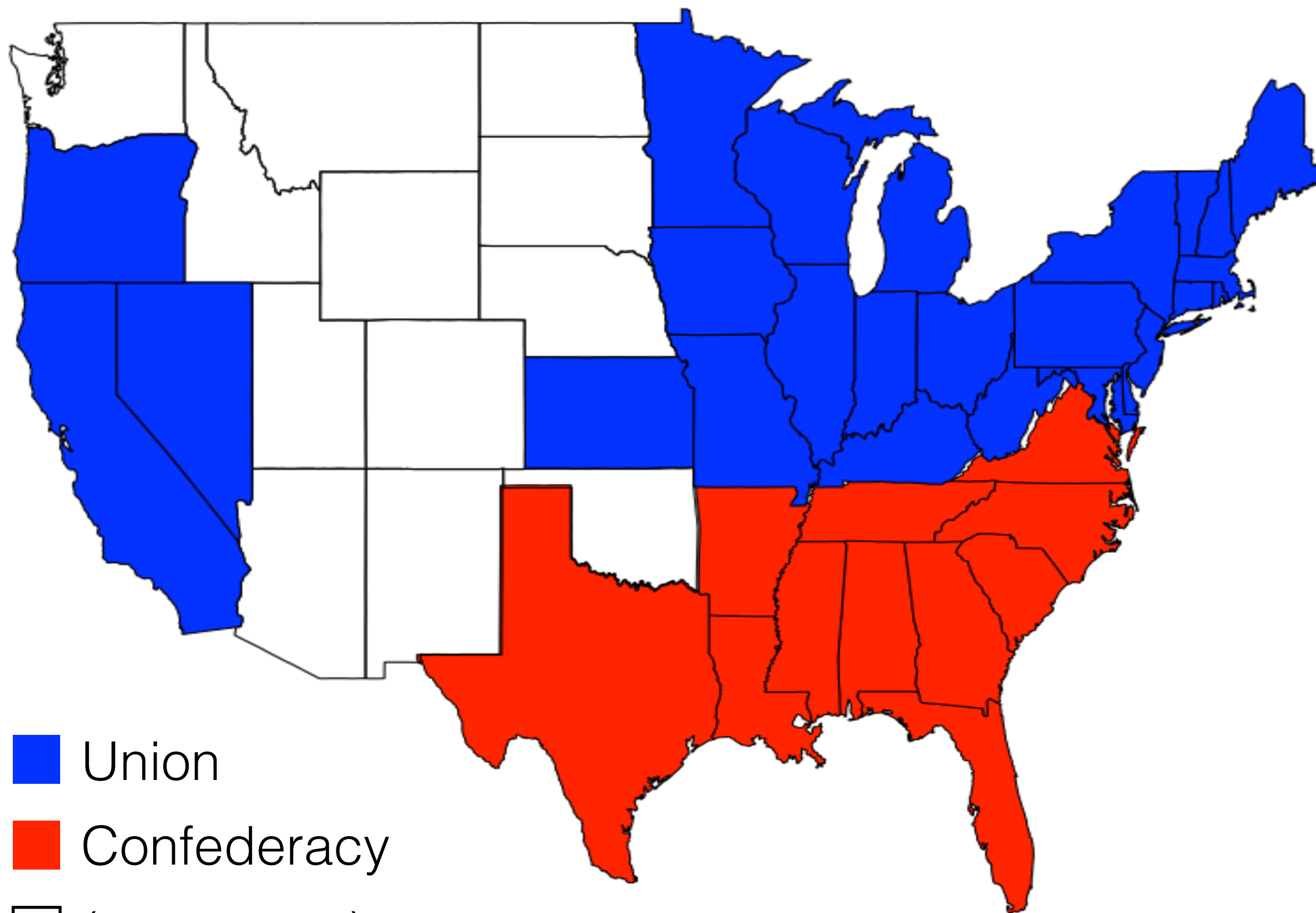
1861



1861



1865



Union



Confederacy



(not a state)


# **Dataset:** United States Presidential Elections

Incumbent Party: Do Democrats or Republicans get a higher percentage of the vote when they are an incumbent?

Civil War Country: Do Union or Confederate states vote differently for incumbents?

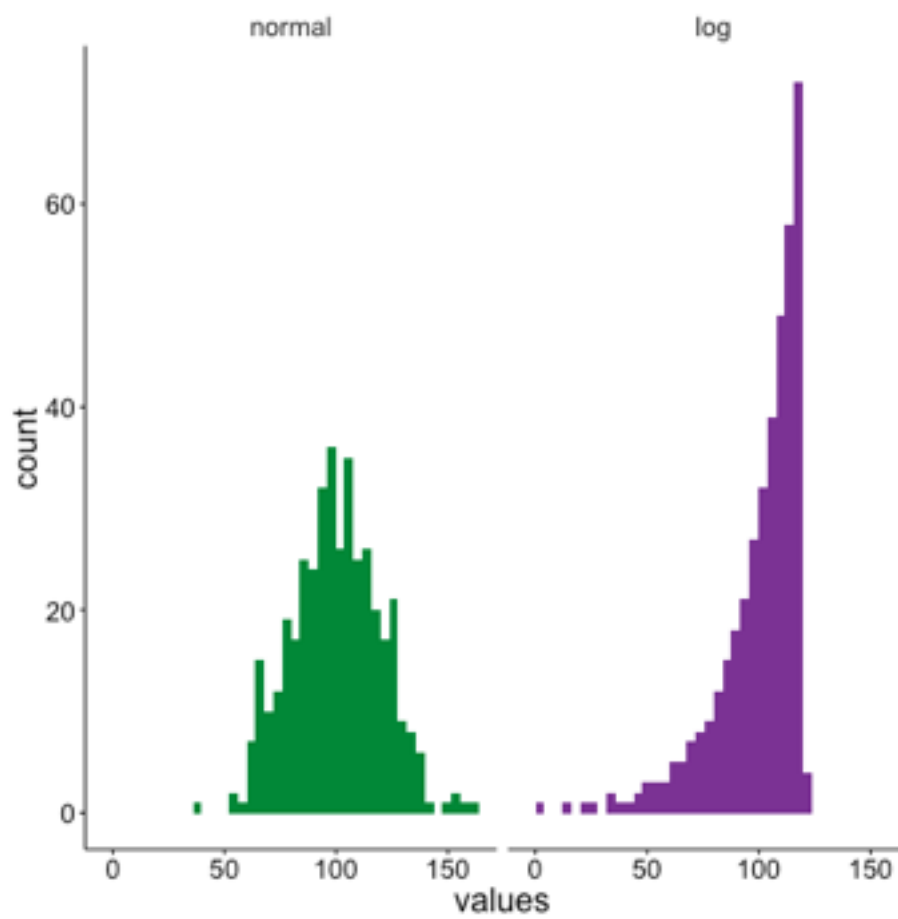
Incumbent Party x Civil War Country: Is there an interaction between these variables?

y	=	percentage of incumbent votes	
x1	=	incumbent party	within
x2	=	civil war country	between
error	=	state	
averaged	=	election year	

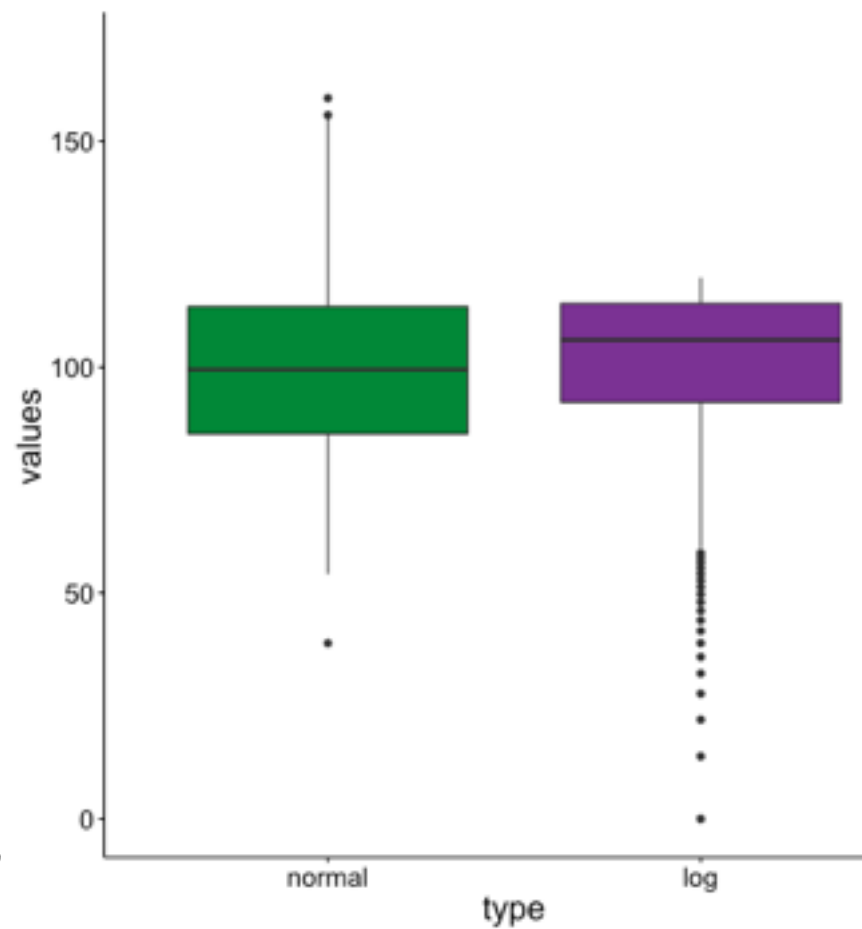


*source: The American Presidency Project*

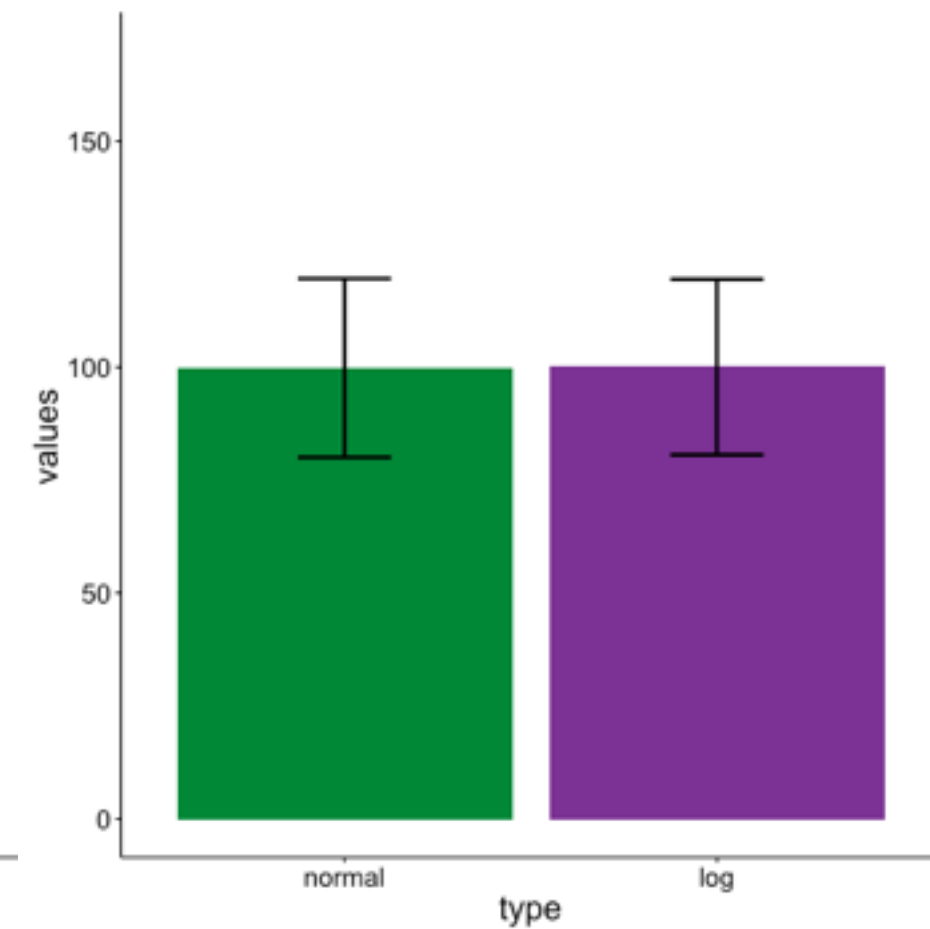
Clearly different...



also pretty  
clearly different...



the same!



## ggplot2

```
incumbent_barplot.plot =  
  ggplot(data_figs_sum,  
    aes(x = civil_war, y = mean,  
      fill = incumbent_party)) +
```

## ggplot2

```
incumbent_barplot.plot =  
  ggplot(data_figs_sum,  
    aes(x = civil_war, y = mean,  
      fill = incumbent_party)) +  
  geom_bar(stat = "identity",  
    position = "dodge")
```

## ggplot2

```
incumbent_barplot.plot =  
  ggplot(data_figs_sum,  
        aes(x = civil_war, y = mean,  
            fill = incumbent_party)) +  
  geom_bar(stat = "identity",  
          position = "dodge") +  
  geom_errorbar(  
    )
```

add  
error bars




## ggplot2

```
incumbent_barplot.plot =  
  ggplot(data_figs_sum,  
    aes(x = civil_war, y = mean,  
        fill = incumbent_party)) +  
  geom_bar(stat = "identity",  
    position = "dodge") +  
  geom_errorbar(aes(ymin = se_low  
    )  
  )  
  )
```

add  
error bars

set minimum point

The diagram consists of two red curved arrows. The first arrow starts at the text 'add error bars' and points to the 'geom\_errorbar' function in the code. The second arrow starts at the text 'set minimum point' and points to the 'ymin = se\_low' argument within the 'aes' function of 'geom\_errorbar'.

## ggplot2

```
incumbent_barplot.plot =  
  ggplot(data_figs_sum,  
        aes(x = civil_war, y = mean,  
            fill = incumbent_party)) +  
  geom_bar(stat = "identity",  
          position = "dodge") +  
  geom_errorbar(aes(ymin = se_low,  
                   ymax = se_high))
```

set minimum point

set maximum point

add

error bars

## ggplot2

```
incumbent_barplot.plot =  
  ggplot(data_figs_sum,  
        aes(x = civil_war, y = mean,  
            fill = incumbent_party)) +  
  geom_bar(stat = "identity",  
          position = "dodge") +  
  geom_errorbar(aes(ymin = se_low,  
                   ymax = se_high),  
               width = 0.2
```

)

add  
error bars

set minimum point

set maximum point

set size

## ggplot2

```
incumbent_barplot.plot =  
  ggplot(data_figs_sum,  
        aes(x = civil_war, y = mean,  
            fill = incumbent_party)) +  
  geom_bar(stat = "identity",  
          position = "dodge") +  
  geom_errorbar(aes(ymin = se_low,  
                  ymax = se_high),  
               width = 0.2,  
               position =  
                 position_dodge(0.9))
```

add error bars

set minimum point

set maximum point

set size

set position

**tidyr, dplyr**

```
data_union_stats = data_stats
```

**tidyr, dplyr**

```
data_union_stats = data_stats %>%  
  filter(civil_war ==  
         "union")
```

**tidyr, dplyr**

```
data_union_stats = data_stats %>%  
  filter(civil_war ==  
         "union") %>%  
  spread(  
    )
```

verb



**tidyr, dplyr**

```
data_union_stats = data_stats %>%  
  filter(civil_war ==  
         "union") %>%  
  spread(incumbent_party,  
         )
```

verb



variable to be  
spread out





## **tidyr, dplyr**

```
data_union_stats = data_stats %>%  
  filter(civil_war ==  
          "union") %>%  
  spread(incumbent_party,  
         perc_incumbent_mean)
```

verb

variable to be  
spread out

variable to put in  
new spread out variables

## tidyr, dplyr

```
data_union_stats = data_stats %>%  
  filter(civil_war ==  
         "union") %>%  
  spread(incumbent_party,  
         perc_incumbent_mean)
```

state	incumbent_ party	perc_incumbent_ _mean
Connecticut	democrat	54.30
Connecticut	republican	49.75
Delaware	democrat	54.03
Delaware	republican	50.13
...	...	...
Vermont	democrat	56.18
Vermont	republican	47.45



state	democrat	republican
Connecticut	54.30	49.75
Delaware	54.03	50.13
...	...	...
Vermont	56.18	47.45

# End of Lesson Questions

But aren't percentages *really*  
just summarized count data?

But we had to drop a bunch of Union states,  
isn't that a problem?

But Alabama was missing one Democrat data point,  
isn't it not balanced?

But what about the variance for 'year',  
shouldn't we try and account for that too?

**generalized linear mixed effects models**