Statistics with SpaRRows

Lecture 12

Julia Schroeder

Julia.schroeder@imperial.ac.uk

ANalysis **O**f **Va**riance – ANOVA

- ANOVA

- Testing for difference of **variances** between groups

- One **categorical variable as explanatory variable**

ANOVA

- Null-hypothesis: groups are equal

ANOVA

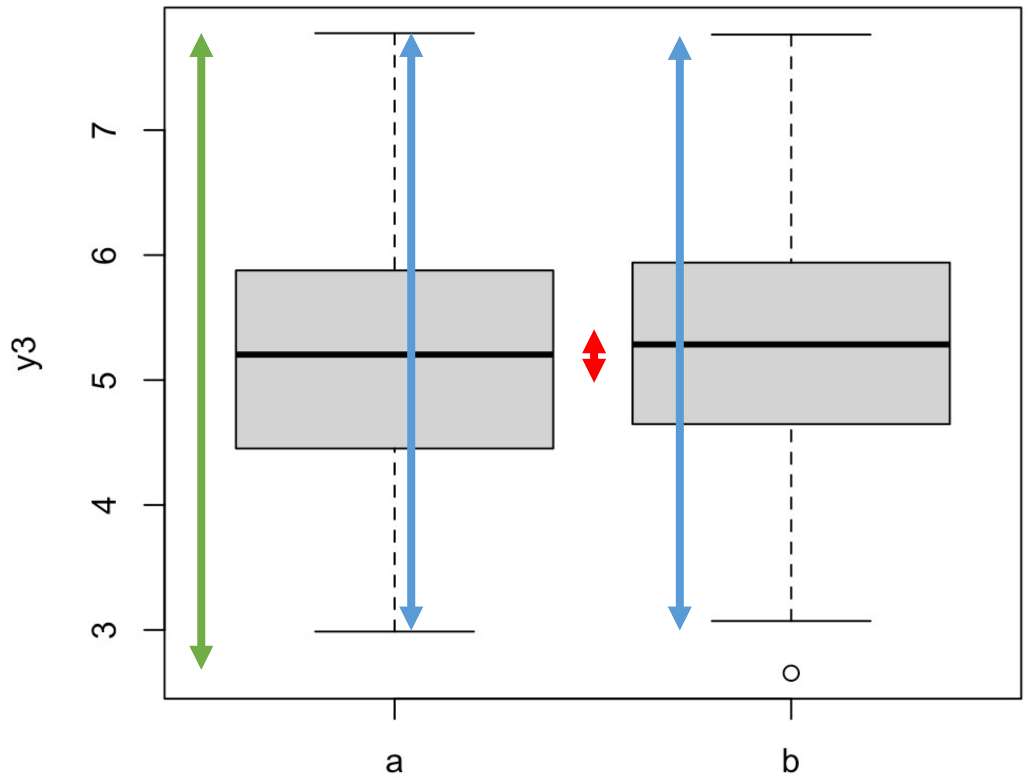
- Null-hypothesis: groups are equal
- Idea: variability among groups is larger than within groups

ANOVA

- Hypothesis: variance among groups is larger than variance within groups

- AG + WG = T

- → variance calculus rules

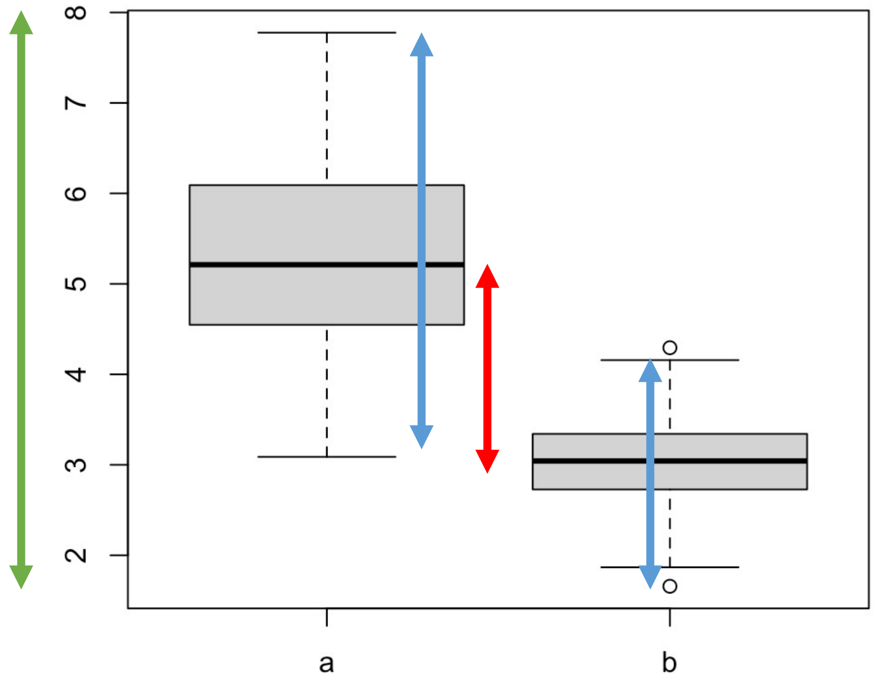


ANOVA

- Hypothesis: variance among groups is smaller than variance within groups

- AG + WG = T

- → variance calculus rules



ANOVA

- So, it’s different from a t-test where we test for a difference in mean

- We test for a difference in variances
- But not between groups – instead, we look at the **ratio** of between group variance to total variance
- From that we can infer that means differ, - but that’s an inference, not the actual test like a linear model or t-test

Ratio between group variance – total variance

- Experiment:
- Outcome variable, 3 treatments:

- Testing for difference between treatments.

Anova

Source	*Sum of Squares*
Group	ASS
Residual (or within group)	WSS
Total	TSS

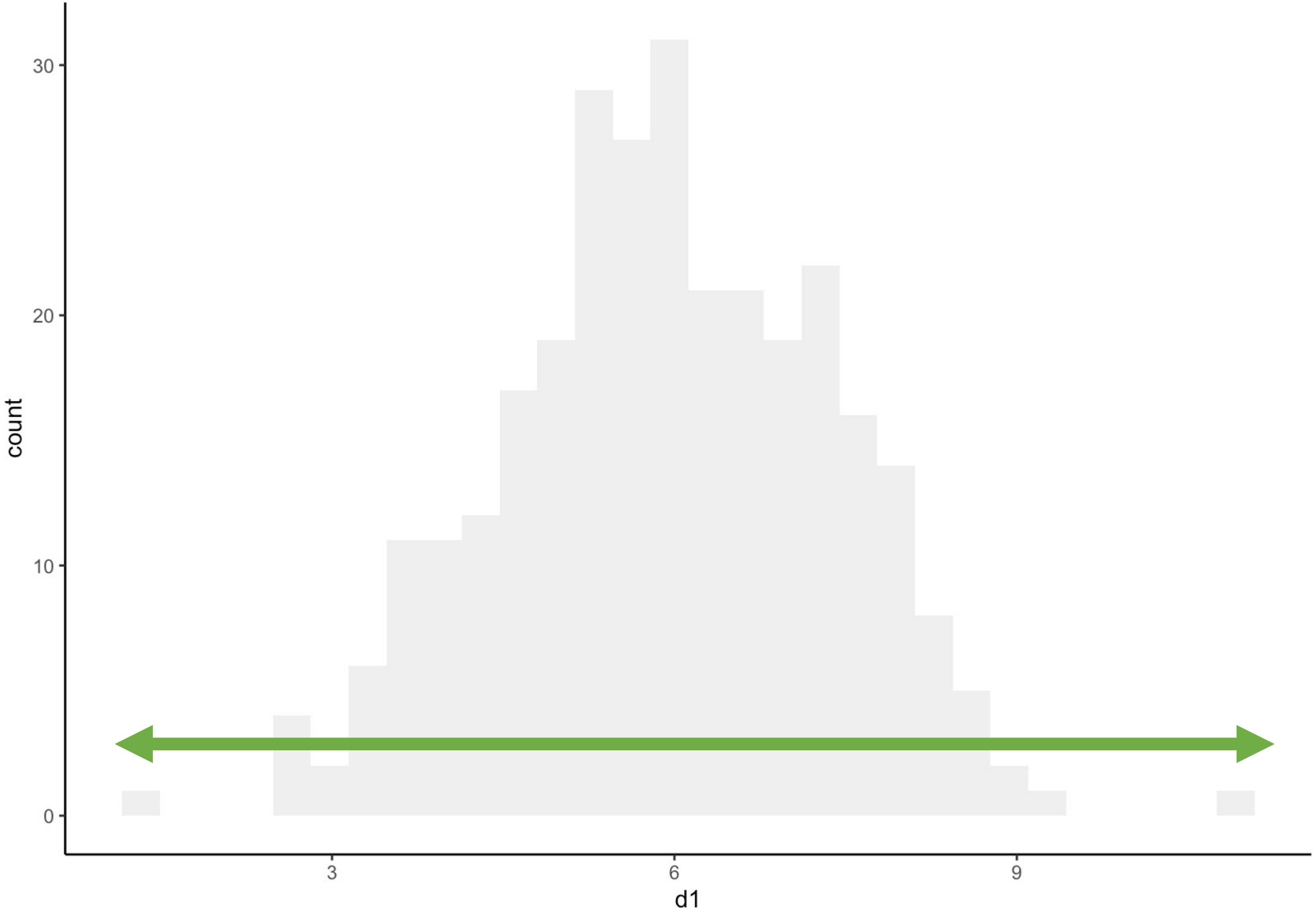
i	y	group
1	1.34	A
2	1.22	A
3	1.27	A
4	1.56	B
5	1.49	B
6	1.52	B
7	1.09	C
8	1.11	C
9	1.08	C

Variances – within and among groups

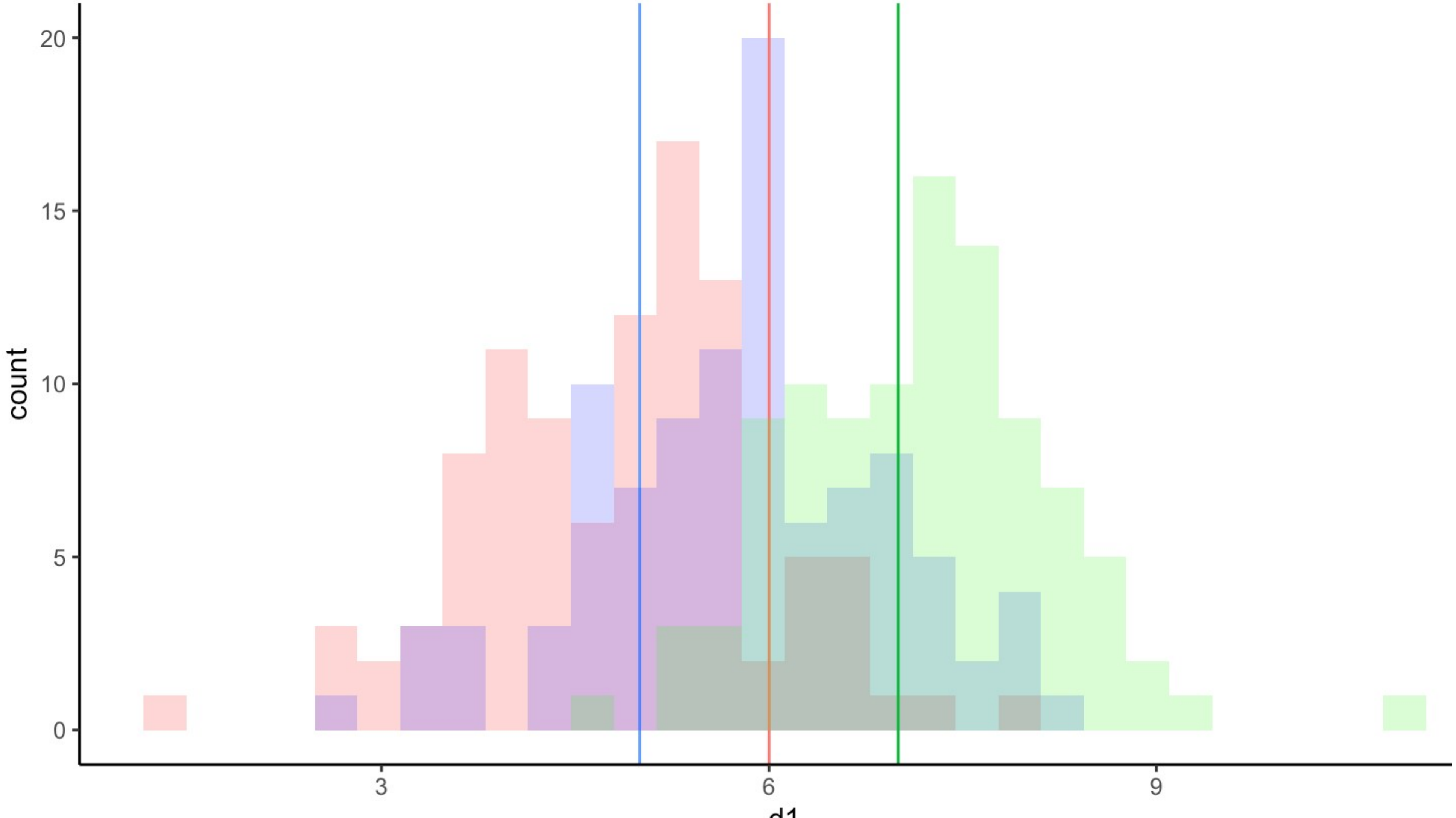
$$\sigma^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

i	y	group
1	1.34	A
2	1.22	A
3	1.27	A
4	1.56	B
5	1.49	B
6	1.52	B
7	1.09	C
8	1.11	C
9	1.08	C

Total sums of squares.



So we have groups..



Notation

efers to data

efers to grand mean

i	y
1	1.34
2	1.22
3	1.27
4	1.56
5	1.49
6	1.52
7	1.09
8	1.11
9	1.08

Notation

efers to data

efers to grand mean

i	y	group
1	1.34	A
2	1.22	A
3	1.27	A
4	1.56	B
5	1.49	B
6	1.52	B
7	1.09	C
8	1.11	C
9	1.08	C

Notation

efers to data

efers to grand mean

efers to data within groups

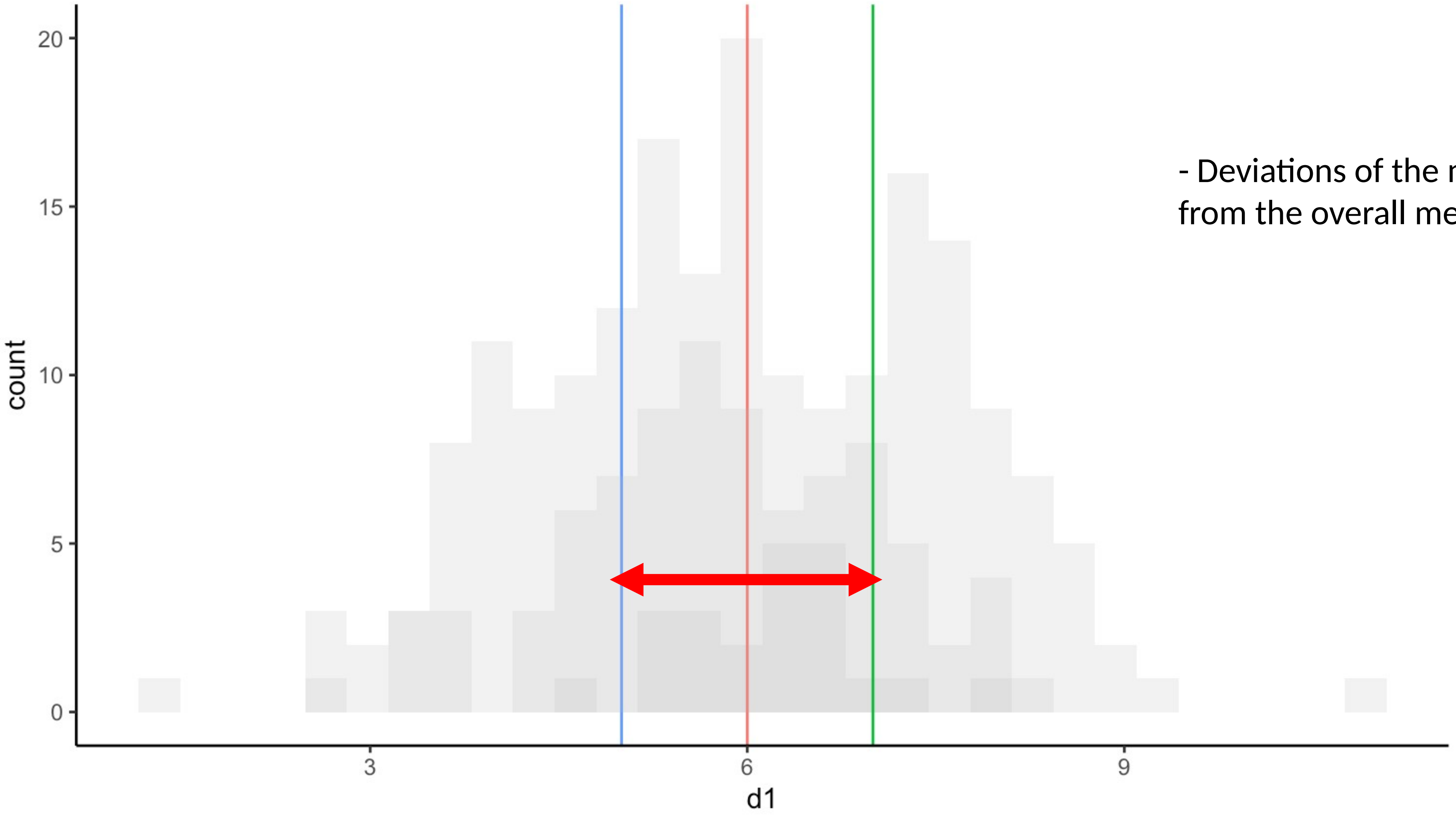
efers to means of each group

efers to grand total

i	y	group	j	I,j
1	1.34	A	1	1,1
2	1.22	A	1	2,1
3	1.27	A	1	3,1
4	1.56	B	2	1,2
5	1.49	B	2	2,2
6	1.52	B	2	3,2
7	1.09	C	3	1,3
8	1.11	C	3	2,3
9	1.08	C	3	3,3

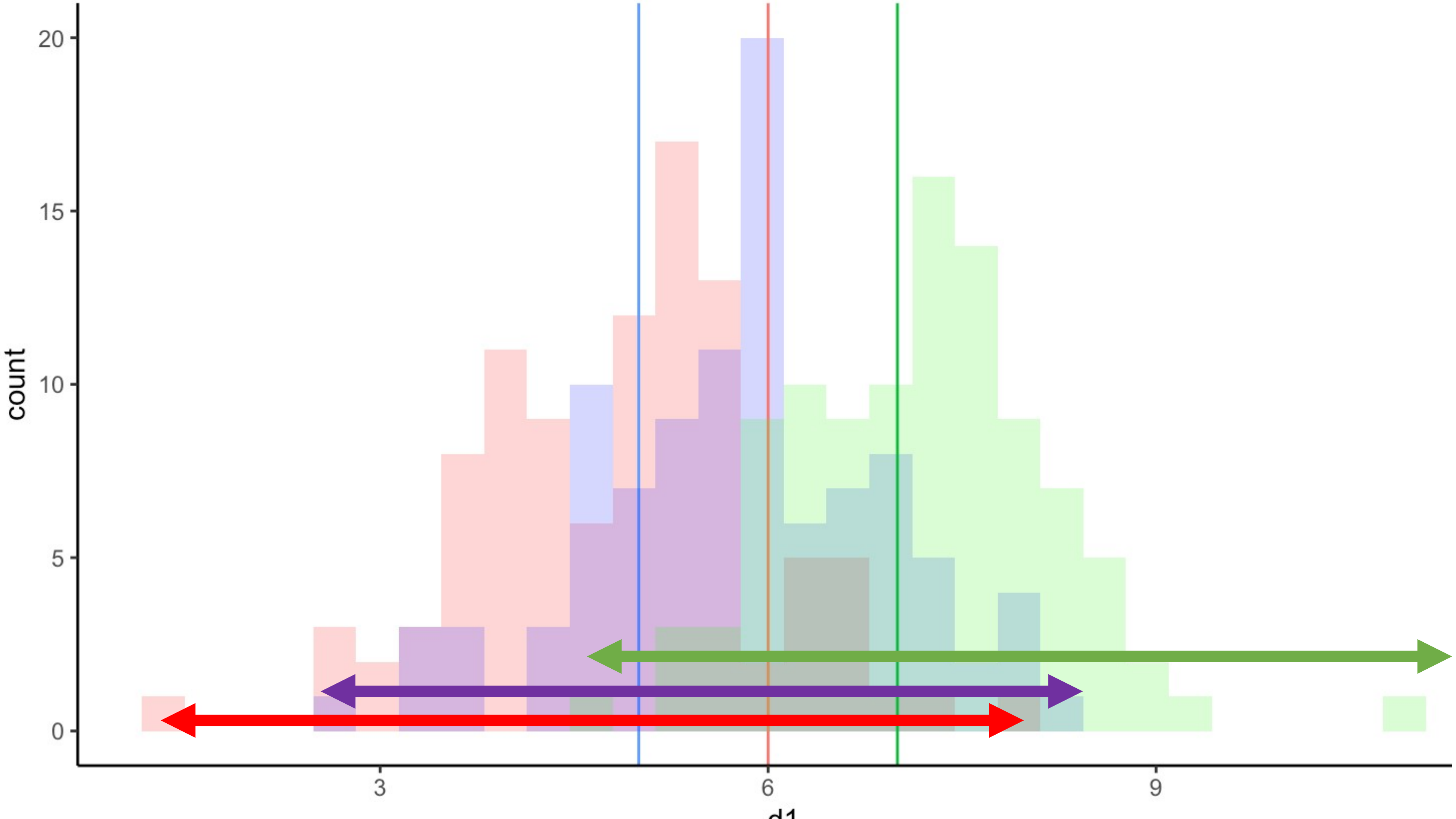
Among-group sum of squares:

=



- Deviations of the means from the overall mean

Within-group sums of squares:



Source	*Sum of Squares*	*d.f.*	*Mean Squares*	*F*
Among-group	ASS	k-1		
Within-group	WSS	n-k		
Total	TSS	n-1		

- Sum of squares among groups: ASS
- Sum of squares within groups = residual: WSS
- Total sum of squares: TSS = ASS + WSS

- k = n of groups (length of j)
- n = sample size (length of i)

Source	*Sum of Squares*	*d.f.*	*Mean Squares*	*F*
Among-group	ASS	k-1		
Within-group	WSS	n-k		
Total	TSS	n-1		

```
> anova(lm(dt$d1~dt$group))
Analysis of Variance Table

Response: dt$d1
           Df Sum Sq Mean Sq F value    Pr(>F)
dt$group    2 268.11 134.055   112.2 < 2.2e-16 ***
Residuals 297 354.85   1.195
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation:

- There are differences between the means
- We don’t know where or how much
- → post-hoc tests to see where

ANOVA:

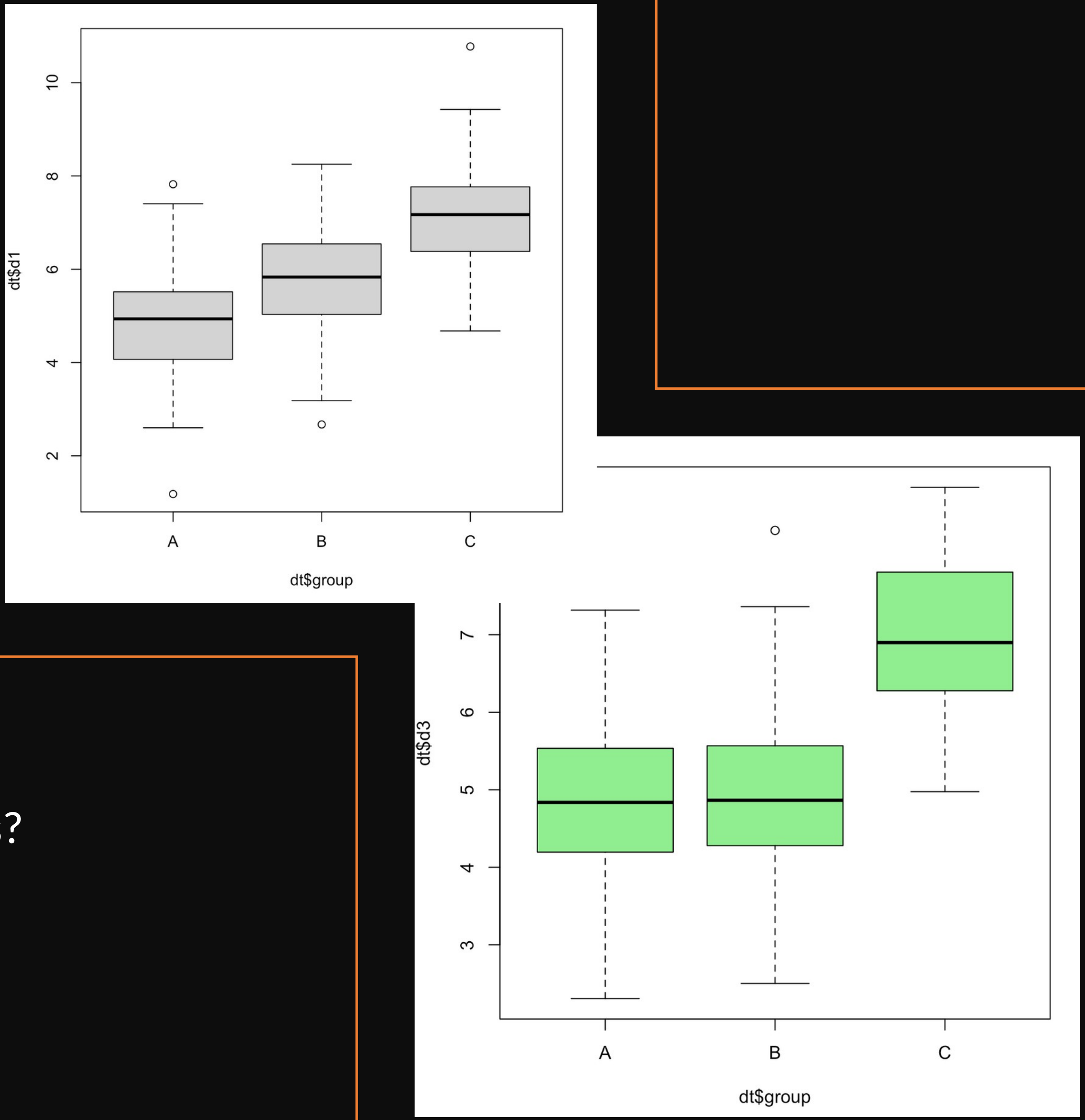
- A continuous response variable
- A factorial explanatory variable with two or more levels
- Provides secondary statistics only
- Needs post-hoc testing for effect sizes and to find out which categories differ

Linear model

- A continuous response variable
- A continuous explanatory variable AND/OR a factorial explanatory variable
- Provides primary statics
- We get effect sizes (means) and precision of each level, thus can judge differences

Battle: ANOVA vs linear model summary

Which one is more useful to describe the two results?



```
> anova(lm(dt$d1~dt$group))
Analysis of Variance Table

Response: dt$d1
           Df Sum Sq Mean Sq F value    Pr(>F)    
dt$group    2 268.11 134.055   112.2 < 2.2e-16 ***
Residuals 297 354.85   1.195                      
---
```

```
> anova(lm(dt$d3~dt$group))
Analysis of Variance Table

Response: dt$d3
           Df Sum Sq Mean Sq F value    Pr(>F)    
dt$group    2 303.03 151.515  156.78 < 2.2e-16 ***
Residuals 297 287.02   0.966                      
---
```

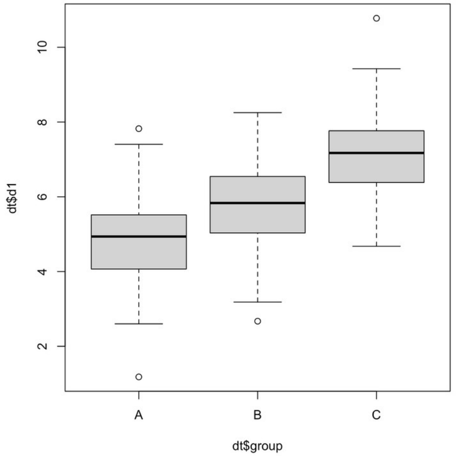
```
> summary(lm(dt$d1~dt$group))

Call:
lm(formula = dt$d1 ~ dt$group)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6782 -0.7535  0.0549  0.6866  3.6178 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.8589     0.1093  44.453  < 2e-16 ***
dt$groupB     0.9113     0.1546   5.895 1.01e-08 ***
dt$groupC     2.2992     0.1546  14.874  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.093 on 297 degrees of freedom
Multiple R-squared:  0.4304,	Adjusted R-squared:  0.4265 
F-statistic: 112.2 on 2 and 297 DF,  p-value: < 2.2e-16
```



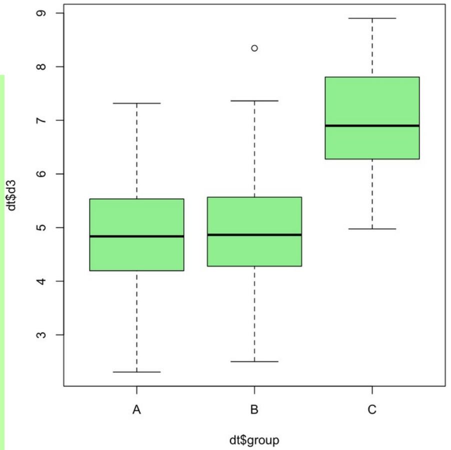

```
> summary(lm(dt$d3~dt$group))

Call:
lm(formula = dt$d3 ~ dt$group)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5761 -0.7155 -0.0378  0.7630  3.4675 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.882305   0.098306  49.664   <2e-16 ***
dt$groupB   -0.004615   0.139025  -0.033    0.974    
dt$groupC    2.129697   0.139025  15.319   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9831 on 297 degrees of freedom
Multiple R-squared:  0.5136,	Adjusted R-squared:  0.5103 
F-statistic: 156.8 on 2 and 297 DF,  p-value: < 2.2e-16
```



How to get anova table in R?

```
> anova(lm(dt$d1~dt$group))
Analysis of Variance Table

Response: dt$d1
           Df Sum Sq Mean Sq F value    Pr(>F)
dt$group    2 268.11 134.055   112.2 < 2.2e-16 ***
Residuals 297 354.85   1.195
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Anova vs linear model summary:

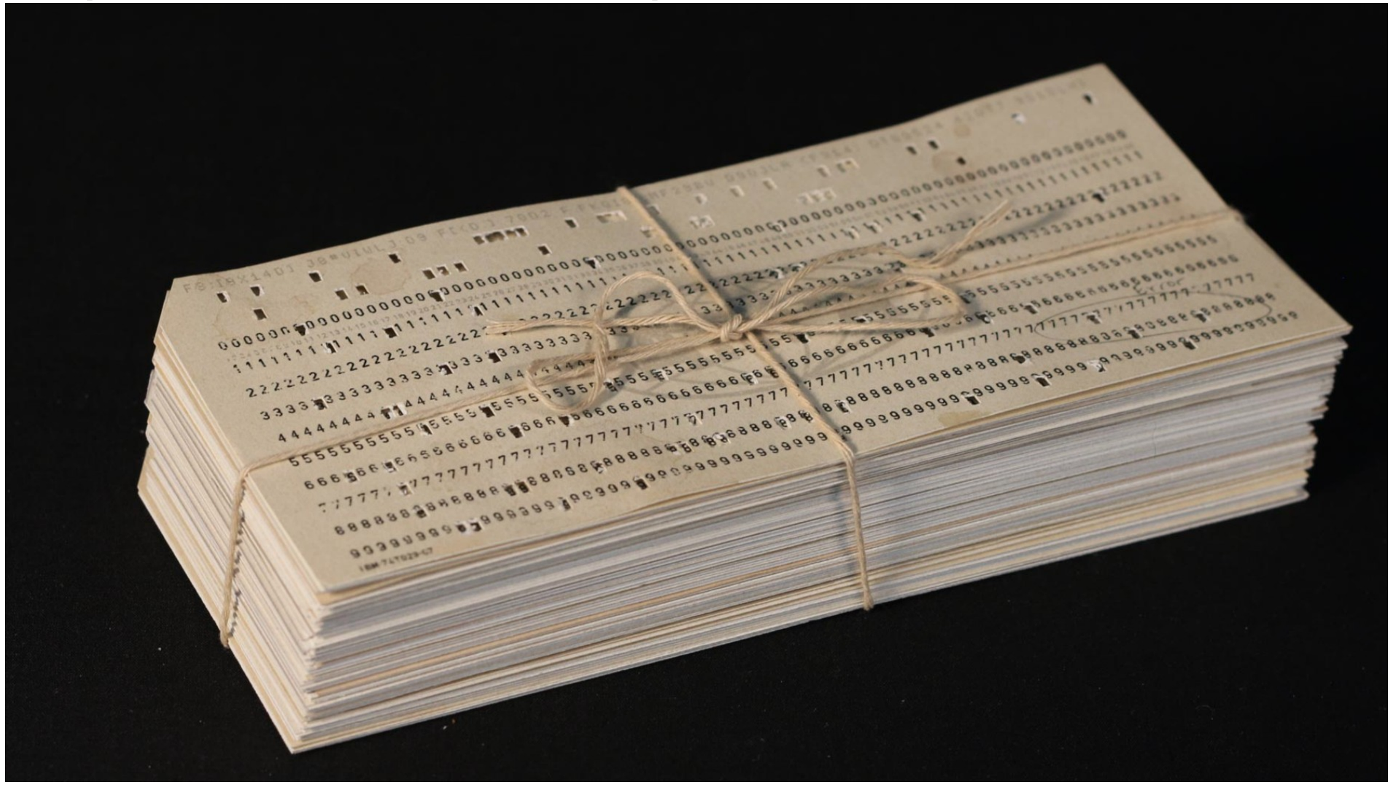
- ANOVA is calculated from results of a linear model in R
- ANOVA *is* a linear model, just a substandard way to report it
- Because we only get secondary statistics (F)
- MSSs provide information about variance (somewhat primary)

- Linear model summary provides primary statiscs
- And F-statistics
- But information about variance is missing (but residual descriptives)

So, why the hype about ANOVA?

- MSs are less computationally intensive calculate than model estimates

- 40-50 years ago you'd either do your calculations by hand (with a pocket calculator)
- Or use hole punch cards

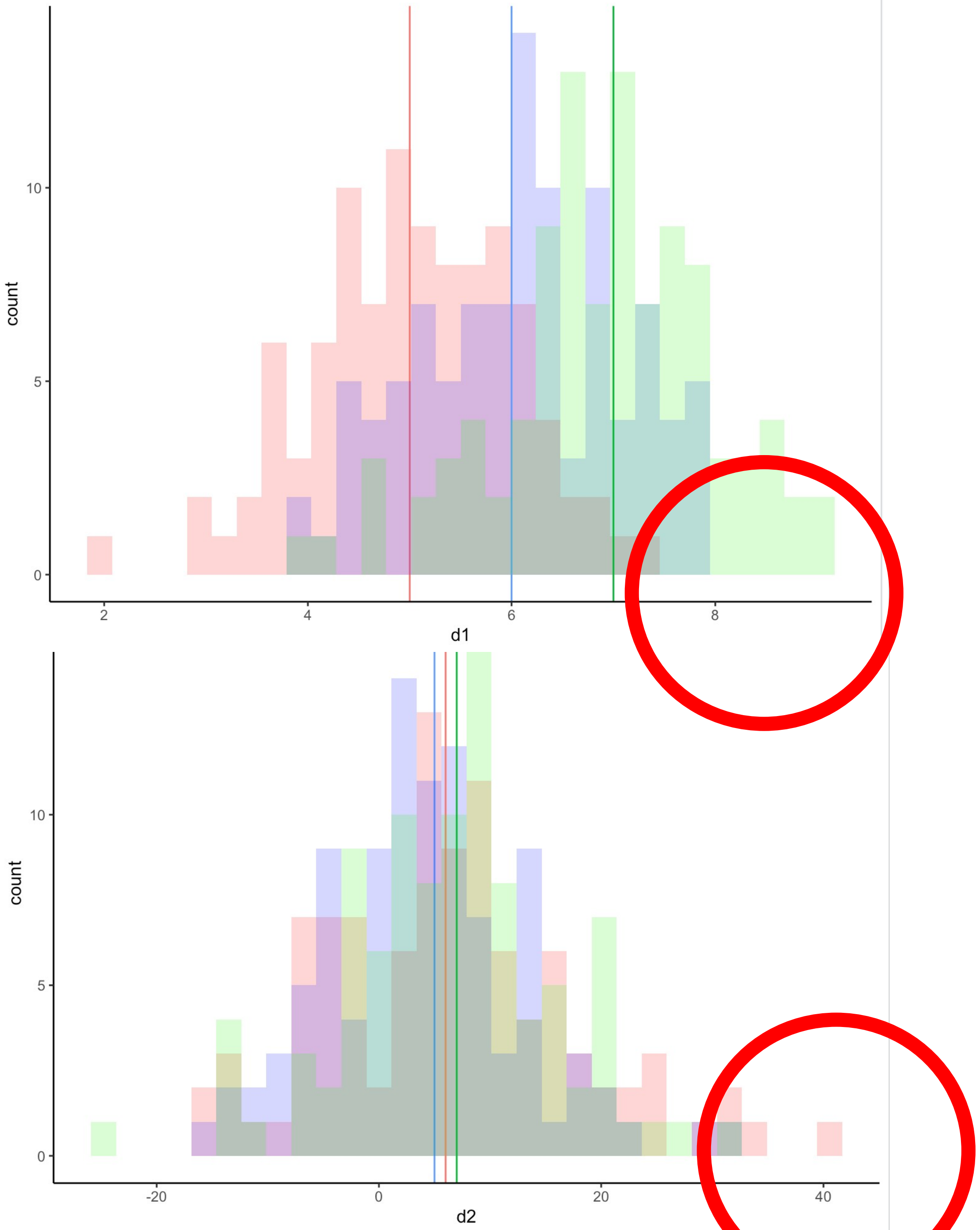


https://www.youtube.com/watch?v=KG2M4ttzBnY

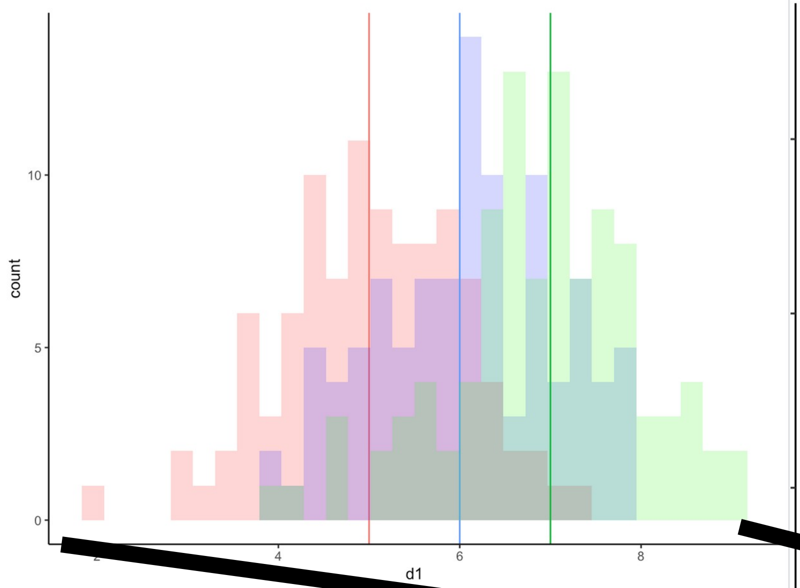
- Need to punch data and code

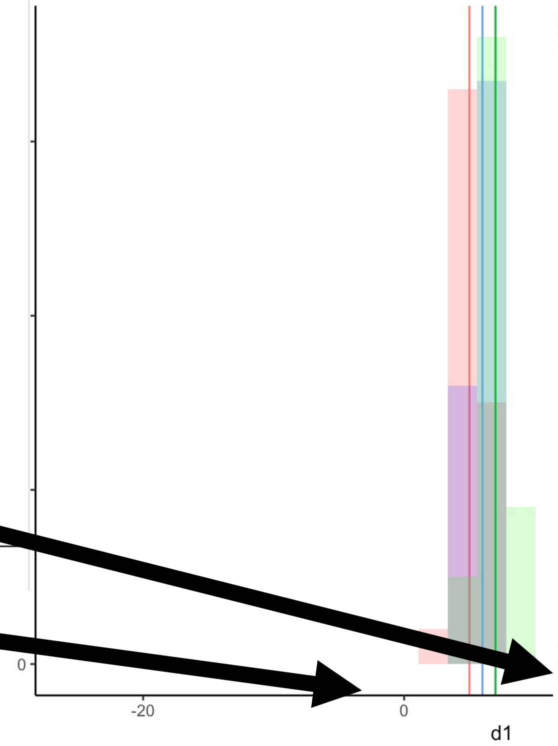
- Model estimates more computationally intensive (weeks!)
- More punching – more room for error
- So, an iterative approach was useful – if not significant no need to put in the extra work
- ANOVA today is more of a relic, habit, tradition
- Use linear models instead!

ANOVA is not really about means, it is an analysis of variances:



Same means
Different variances





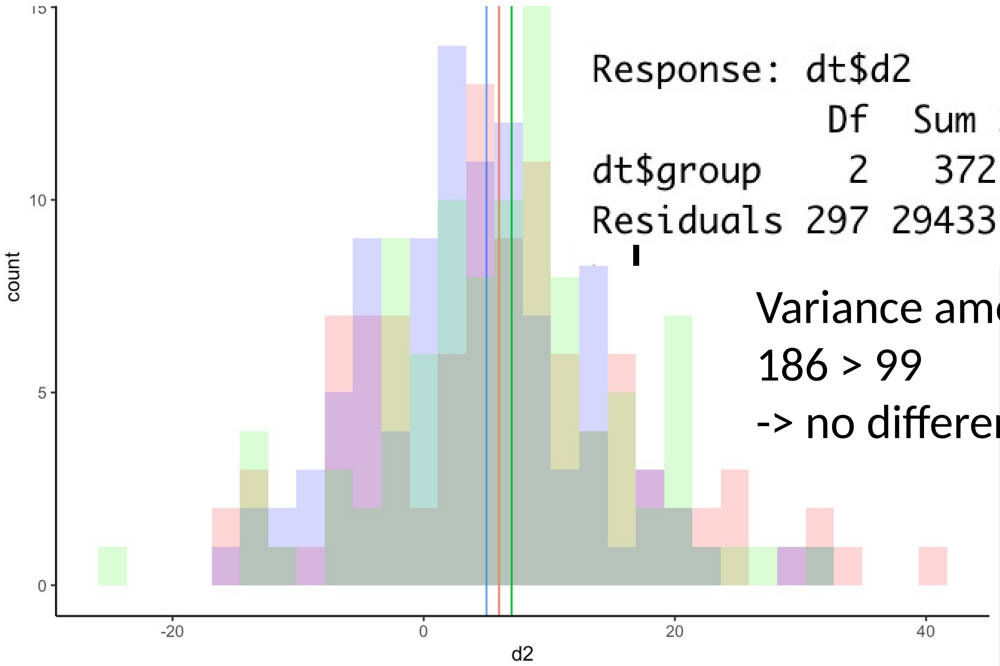

```
> anova(lm(dt$d1~dt$group))
Analysis of Variance Table

Response: dt$d1
           Df Sum Sq Mean Sq F value    Pr(>F)
dt$group    2 178.81  89.403  86.754 < 2.2e-16 ***
Residuals 297 306.07   1.031
```

Variance between groups < variance within groups:
89 > 1
-> differences between groups

```
> anova(lm(dt$d2~dt$group))
Analysis of Variance Table

Response: dt$d2
           Df  Sum Sq Mean Sq F value Pr(>F)
dt$group    2   372.1 186.032  1.8771 0.1548
Residuals 297 29433.9  99.104
```



Variance among groups > variance within groups:
186 > 99
-> no differences between groups

But what if we want to know about variances ***and* mean estimates?**

- Linear mixed models estimate variance components and fixed effects simultaneously!

Among group variance

Within group variance

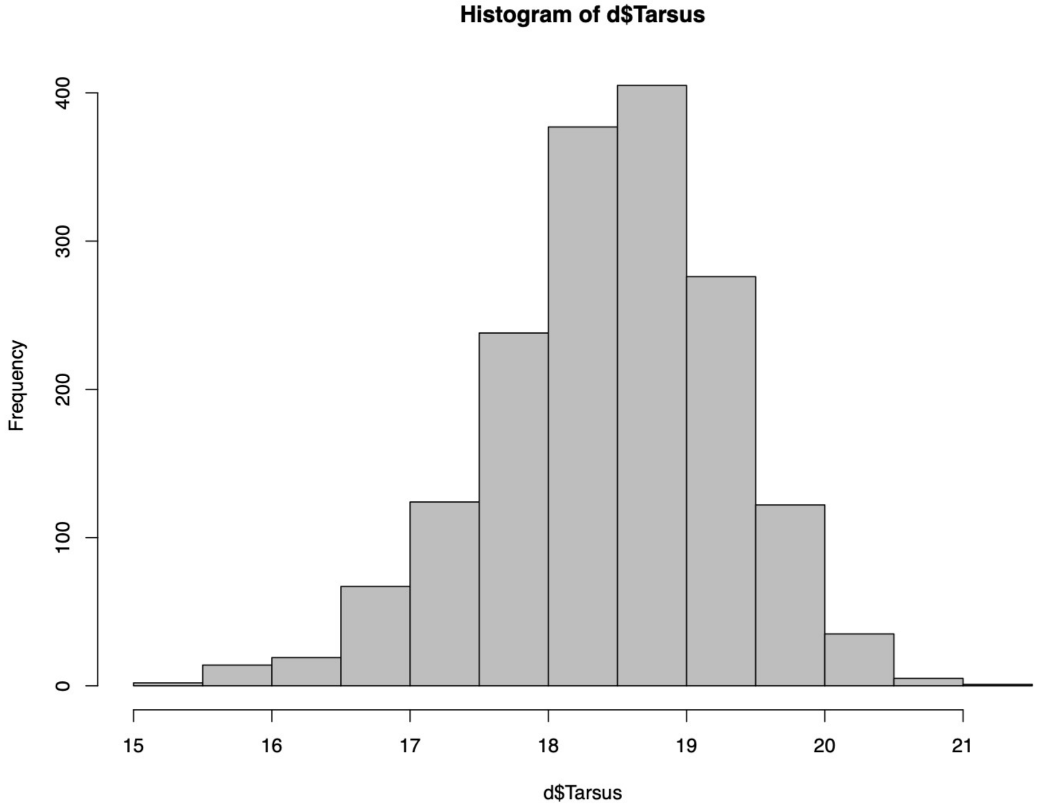
Total variance:
Va + Vw = Vt
=3.9

groupC 2.2992 2.3286 0.987

We've come full circle:

- Linear mixed models estimate variance components and fixed effects simultaneously

- Linear mixed models estimate **spread** and **centrality simultaneously**



Describe data:

- Centrality
- Spread