Analysis of Variance (ANOVA)

The t-test for adults.

Use ANOVA when...

- you have a numeric response (DV)
- and a categorical IV with two or more levels
  - if there are only 2 levels, the t-test is better
  - ANOVA is always nondirectional (i.e., it does not test hypotheses like "greater than" or "less than" but only hypotheses like "different than")
    - null hypothesis - all sampled populations have the same mean
    - alternative - at least one of the means is different
- you can also use it when you have two or more IVs
- very complex experimental designs can be tested with ANOVA - it is probably the most frequently used statistical technique in psychology

Experimental Design

```
> CA = read.csv("http://ww2.coastal.edu/kingw/psyc480/data/caffeine.csv")
> CA
       group dose tapping
1    control    0     242
2    control    0     245
3    control    0     244
4    control    0     248
5    control    0     247
6    control    0     248
7    control    0     242
8    control    0     244
9    control    0     246
10   control    0     242
11  low.dose  100     248
12  low.dose  100     246
13  low.dose  100     245
14  low.dose  100     247
15  low.dose  100     248
16  low.dose  100     250
17  low.dose  100     247
18  low.dose  100     246
19  low.dose  100     243
20  low.dose  100     244
21 high.dose  200     246
22 high.dose  200     248
23 high.dose  200     250
24 high.dose  200     252
25 high.dose  200     248
26 high.dose  200     250
27 high.dose  200     246
28 high.dose  200     248
29 high.dose  200     245
30 high.dose  200     250
```

Use this command to retrieve the data in R if you want to look at it.

Caffeine and finger tapping data (from Hand, et al.)

* the subjects were given a dose of caffeine and then asked to tap a keypad as rapidly as they could for one minute - it's a common neurological test of the integrity of the motor system
* IV is dose of caffeine (we will use the variable labeled "group")
* DV is finger taps/min (tapping)
* three independent groups (practiced male subjects)
* this is called a between-groups or between-subjects design (each subject was used in only one of the conditions and subjects were not paired or matched)

Special Note

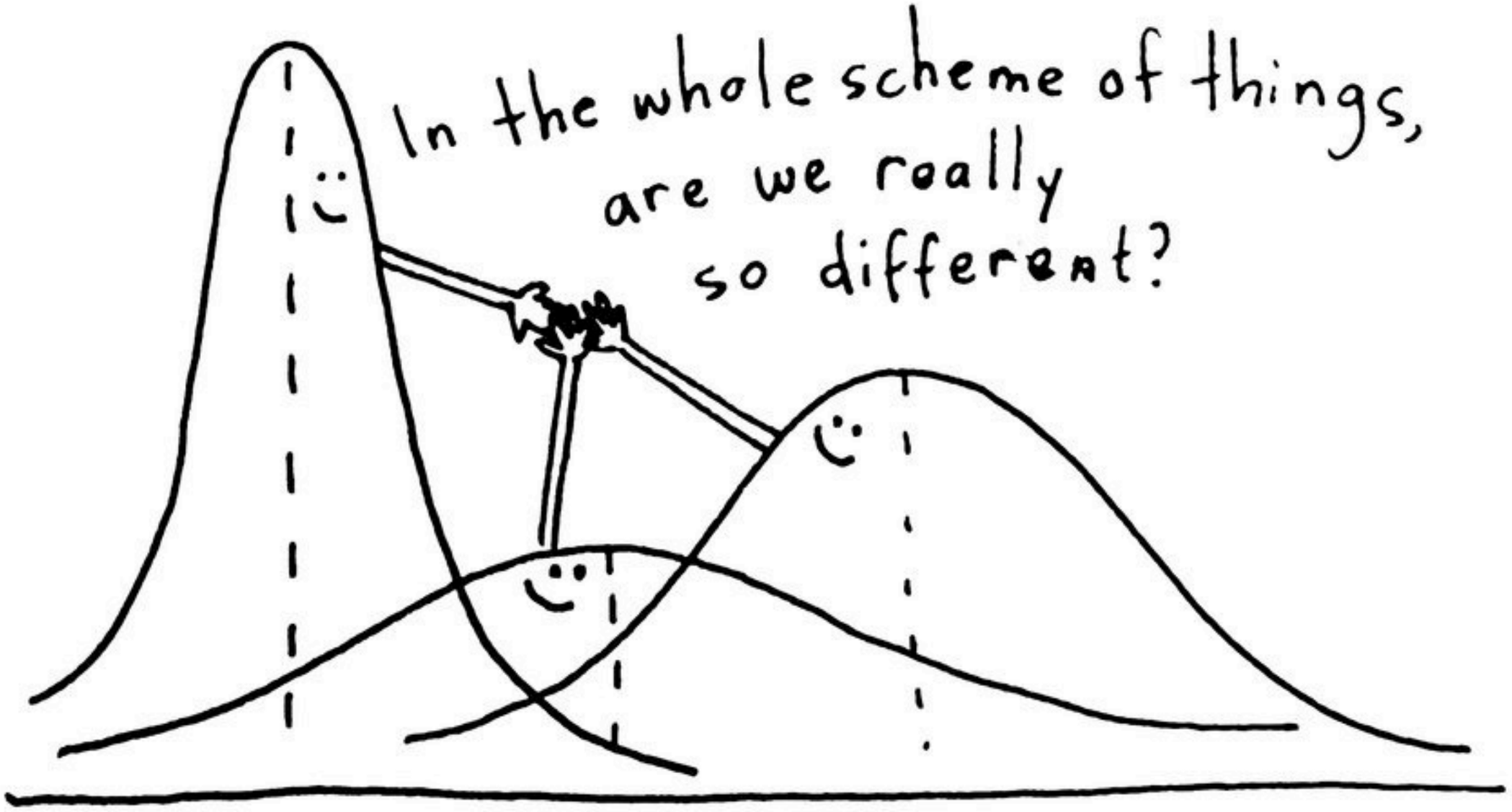
```
group,dose,tapping
control,0,242
control,0,245
control,0,244
control,0,248
control,0,247
control,0,248
control,0,242
control,0,244
control,0,246
control,0,242
low.dose,100,248
low.dose,100,246
low.dose,100,245
low.dose,100,247
low.dose,100,248
low.dose,100,250
low.dose,100,247
low.dose,100,246
low.dose,100,243
low.dose,100,244
high.dose,200,246
high.dose,200,248
high.dose,200,250
high.dose,200,252
high.dose,200,248
high.dose,200,250
high.dose,200,246
high.dose,200,248
high.dose,200,245
high.dose,200,250
```

This is a proper csv file.

* "comma separated values"
* it is a spreadsheet file saved as plain text (csv)
* or it can be typed into a text editor
* or it can be created in an R script window
* cases (subjects) are in the rows and variables are in the columns - notice there is a column for the IV and a column for the DV (in this case, there is also a "dose" column, which we are not using)
* no spaces (although R doesn't seem to care as long as the commas are there and immediately after the data value)
* R and most other statistical software can read these - in R it is done with the read.csv() command
* R will read it in as a data format called a data frame

What's ANOVA All About?

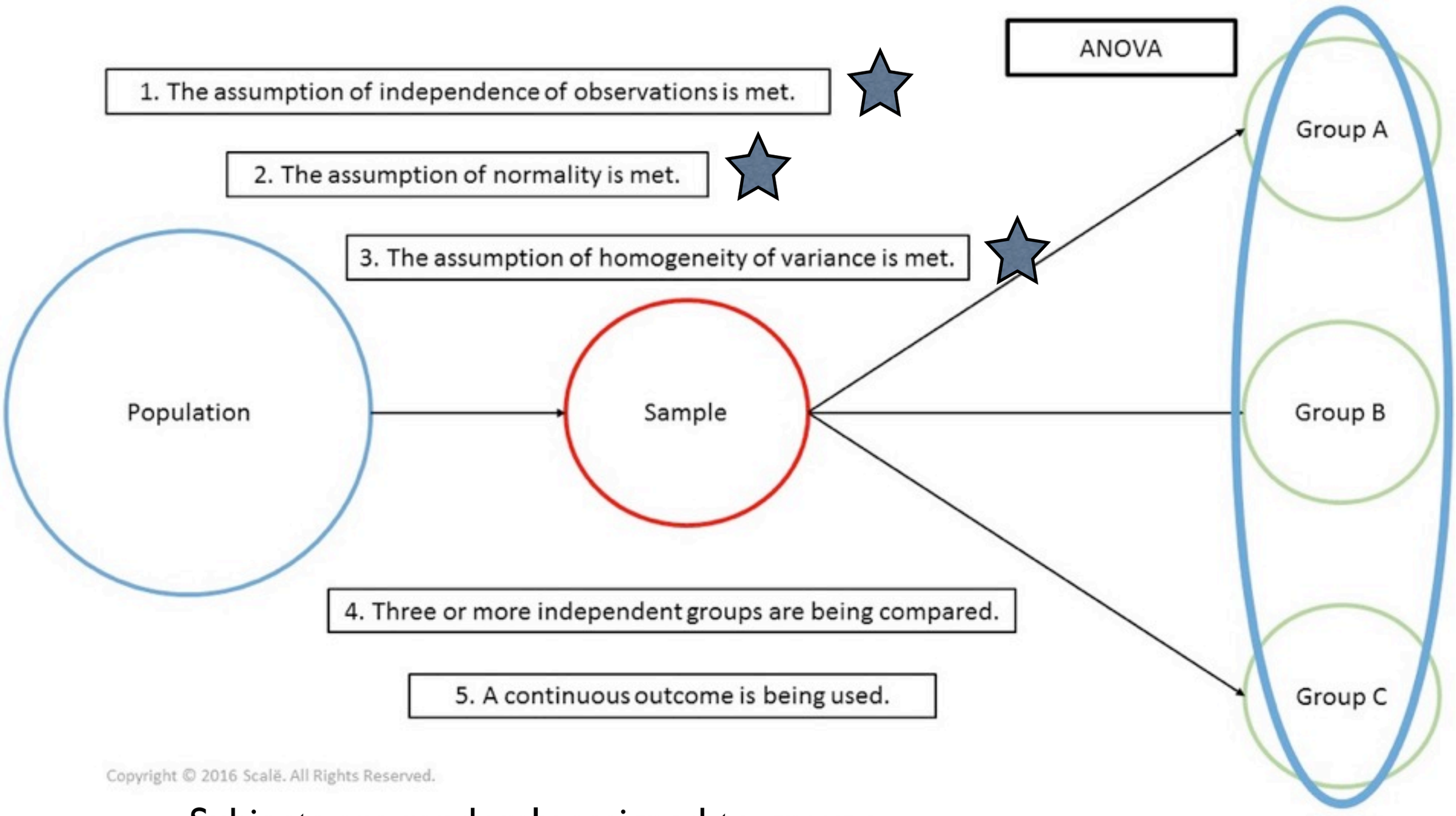
- compares between-groups variability ("explained" variability) to within-groups variability (unexplained variability)
- if there is nothing but random, unexplained variability present, then we expect the between-groups variance to be about the same as the within-groups variance (i.e., all between-groups variance will be due to random error or noise)
- if an effect is present, we expect the between-groups variance to be larger
- just because the means are different doesn't mean an effect is present - why else might the means be different (2 more reasons)?



To put it another way...

* the scores (values of the DV) are (probably) going to overlap across groups
* are the group means different enough that we can conclude with confidence that these really are different groups
* i.e., sampled from populations with different means
* two cases must be considered
    * true (designed, randomized) experiments
    * intact groups (sometimes called quasi-experimental designs)

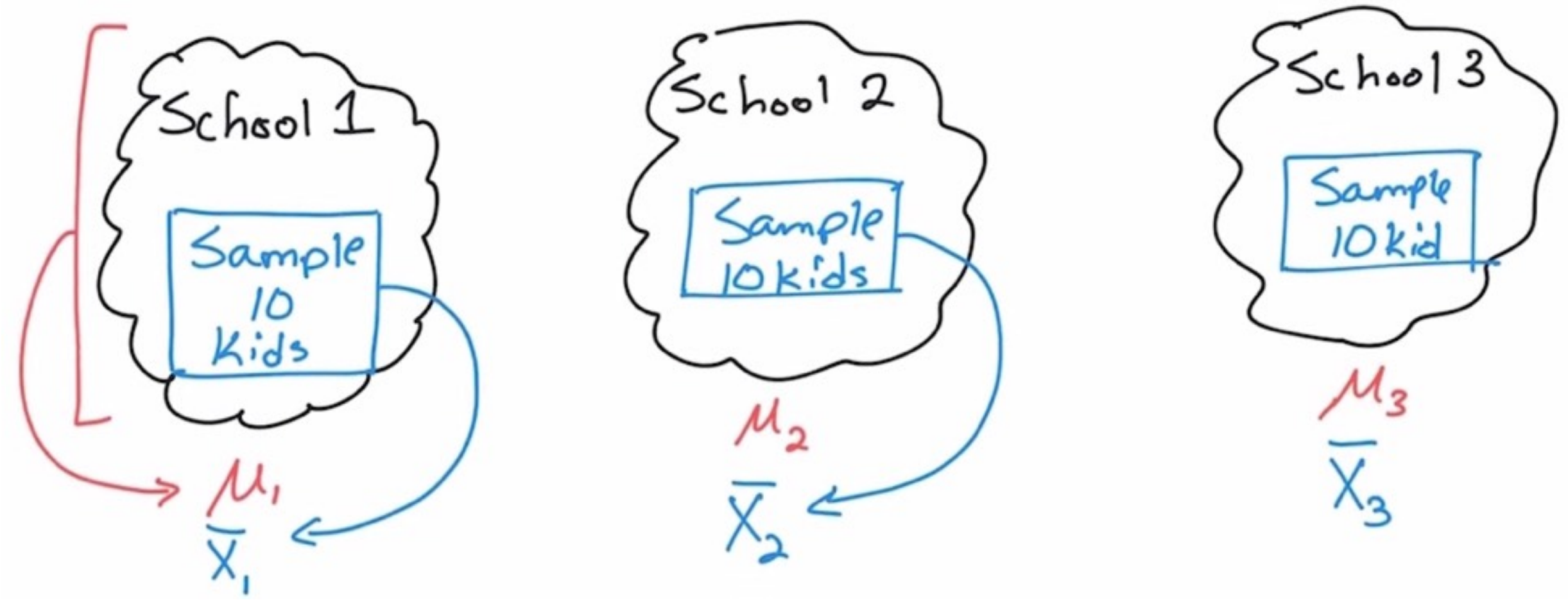
True Experiments



Subjects are randomly assigned to groups.

* true experiment
* did our experimental treatment make the subjects different?
* if so, we can talk about cause and effect (because hopefully the randomization controlled for confounds)

Quasi-experimental Designs (Intact Groups)



We take the subjects as they come (we can't randomly assign kids to schools).

* intact groups (or quasi-experimental design or self-selected subjects)
* were the subjects different to begin with?
* what's the primary problem with this design? (answer: high potential for confounds to be present)
* we cannot talk about cause and effect (well, we can, but we have to be very careful!)

Either way, it's the same calculations.

sum of squares

mean squares - the fancy ANOVA term for variance

Source of Variation	SS	df	MS	F ratio
Between Samples	SSB	k - 1	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Samples	SSW	n - k	$MSW = \frac{SSW}{n-k}$	
Total	SST = SSB+SSW	n - 1		

explained variability

unexplained variability

note: n = N
k = no. of groups

no sum in this box

test statistic:
F is the ratio of two variances

ANOVA Summary Table

SSW = sum of the SSs within the groups = SS.error or SS.within
SST = the SS of the DV values considered as one big group = SS.total
SSB = SST - SSW = SS.treatment or SS.between

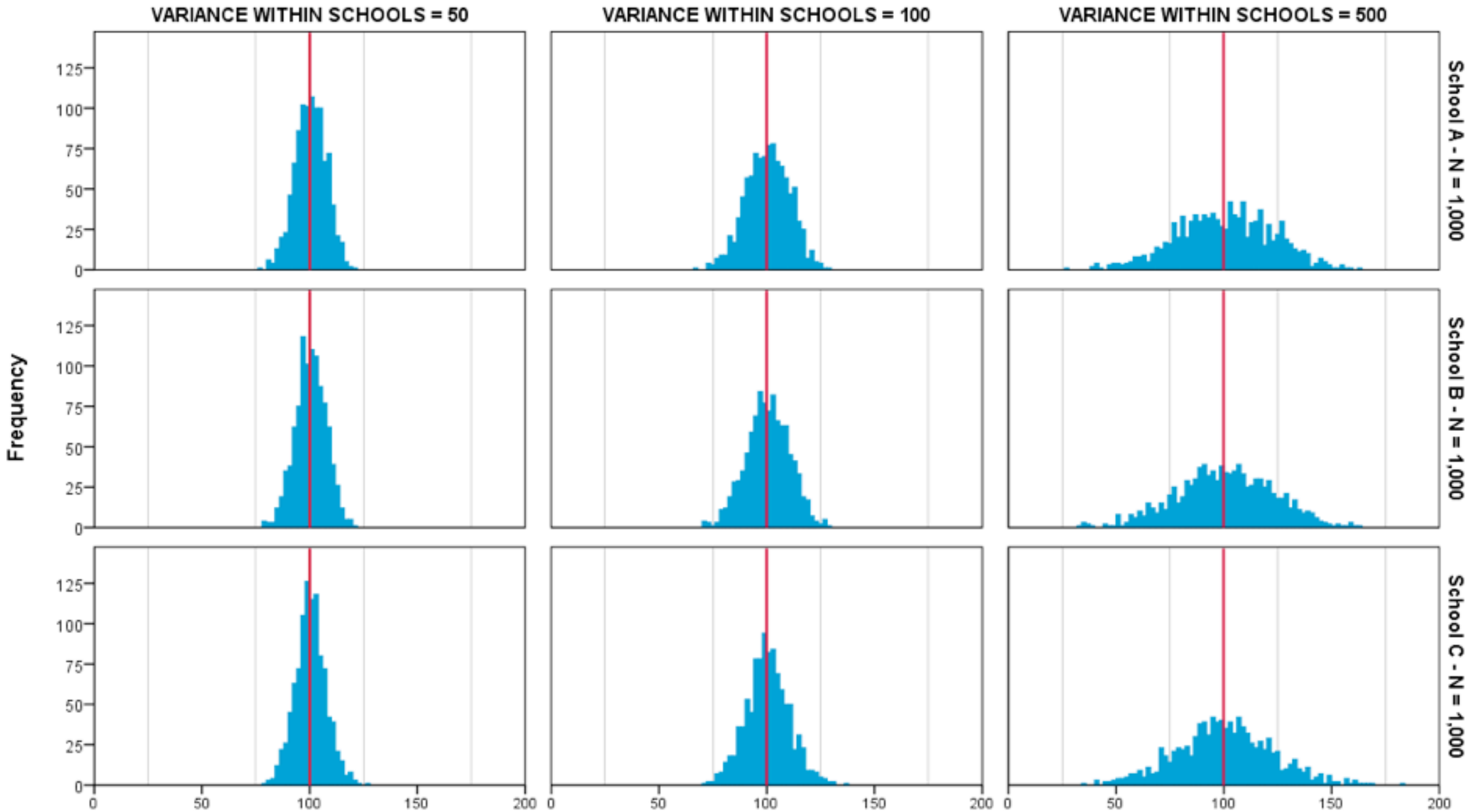
Or you can do it the easy way.

```
> aov.out = aov(tapping ~ group, data=CA)
> summary(aov.out)
             Df Sum Sq Mean Sq F value  Pr(>F)
group         2   61.4  30.700   6.181 0.00616 **
Residuals    27  134.1   4.967
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Know how to read and use this table!

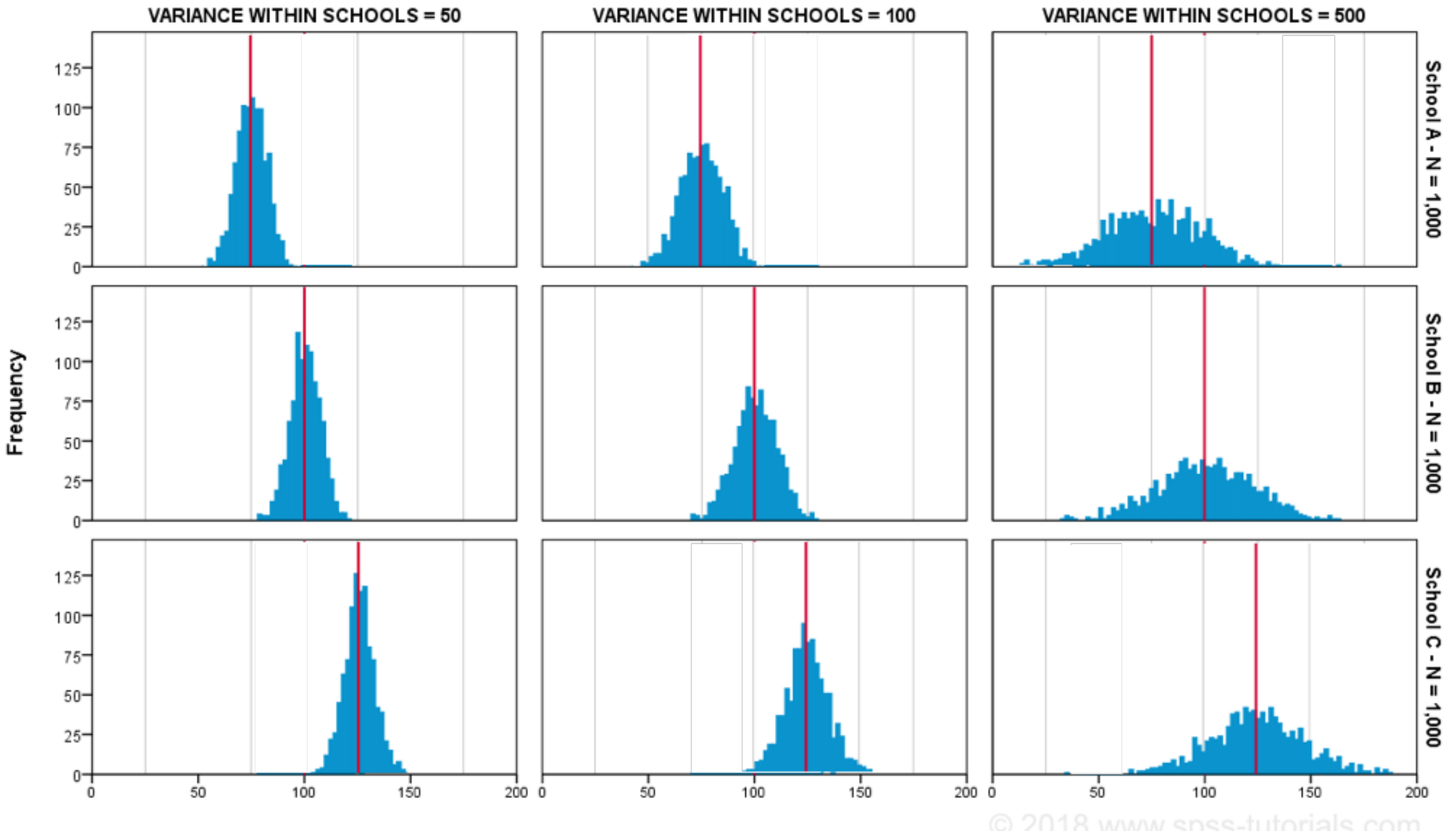
* there is no total line, but dfs and SSes will sum
* the first line is explained variability
* the p-value is under Pr(>F)
* the Residuals line is error or unexplained variability (“residual” = error)
* the asterisks are called significance stars

ONE-WAY ANOVA - SUMS OF SQUARES WITHIN GROUPS



small error moderate error large error

How much between groups variability exists in these three cases? (answer: none)
In which case would it be easiest to see the effect if there were between groups variability?



In which case is it easiest to see (be reasonably confident of) the differences between the groups? (answer: in the case where the groups overlap the least)

Effect Size

- eta-squared = SSB / SST
- explained variability (SSB) divided by total variability (SST)
- 61.4 / (61.4 + 134.1) = 0.314
- interpret this as the proportion of variability explained (PVE)
- caution: do NOT divide by SSW (a common mistake)!
- eta-squared is kind of like a "correlation" squared

Interpreting Eta-squared

(if you have no other way to judge effect size)

- <0.01 is trivial (min. possible value is 0)
- 0.01 - 0.10 is small
- 0.10 - 0.25 is moderate
- 0.25 - 0.50 is large
- above 0.5 is very large (max. possible value is 1.0)

Post Hoc Tests

- After all the above is done, you still don't know which groups are different.
- You just know at least one of them is.
- Post hoc tests tell you which groups are significantly different.
- There are about as many post hoc tests as there are statistics professors who need tenure. (*sarcasm*)

Tukey HSD test

- the "standard" for pairwise comparisons - i.e., groups compared two at a time
- HSD stands for honestly significant difference (because it adjusts the p-value for the number of comparisons being made)
- a very conservative test

```
> TukeyHSD(aov.out)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = tapping ~ group, data = CA)

$group
                     diff        lwr       upr     p adj
high.dose-control     3.5  1.0288609 5.9711391 0.0043753
low.dose-control      1.6 -0.8711391 4.0711391 0.2606999
low.dose-high.dose   -1.9 -4.3711391 0.5711391 0.1562593
```

(Hold it! If A is equal to B and B is equal to C, how can A not be equal to C?)

(You should know how to interpret this table. It gives the difference between the means, a 95% CI for that difference, and a p-value adjusted for the number of comparisons made.)

Fisher LSD test

- basically all pairwise t-tests
- LSD stands for least significant difference
- it does not adjust for the number of comparisons being made
- it probably shouldn't be used if there are more than 3 or 4 comparisons
- it MUST be protected - i.e., should not be done unless the null hypothesis was rejected in the ANOVA

Calculating Fisher LSD

$$t = \frac{mean1 - mean2}{sp * \sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

for df use the full error df from the ANOVA summary

pooled standard deviation - use the square root of MSW for all comparisons (this assumes homogeneity of variance)

(What is a standard deviation?)

Bonferroni test

- also called the Bonferroni-Dunn test
- applies a Bonferroni correction to the Fisher LSD p-values
- multiply the p-values by the number of comparisons you've done
- very conservative

Violating Assumptions

- homogeneity of variance - there is a version of the single-factor ANOVA that does not assume homogeneity, similar to the Welch-corrected t-test

- normality - you can use a nonparametric version of the single-factor ANOVA, such as the Kruskal-Wallis test

  - rank all the data from 1 to N ignoring group membership; tied values are assigned the average of the ranks they would have had if not tied

  - the test statistic H = (N-1) * SSB calculated on the ranks / SSW calculated on the ranks

  - there are tables of critical values of H

Checking Assumptions

- homogeneity of variance - just eyeball the variances
    - if they are "similar enough" then fine
    - there are statistical tests, but they have problems - Levene's test is probably the standard test for homogeneity
- normality - check graphically

A Note on Statistical Significance

- The p-value is a statistical construct and is valid only if all assumptions have been met.
- The alpha level is an arbitrary cutoff. There is no magic in the number .05.
- Statistical significance in no way implies anything about the importance of the result, or even anything about its scientific validity or usefulness.
- It is a decision making aid ONLY.
- It has been widely misinterpreted and misused.
- There is a rebellion against the whole idea.
- Ten years from now, we may not even be talking about statistical significance anymore!