Data Analysis, Statistics, Machine Learning

Leland Wilkinson

Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai

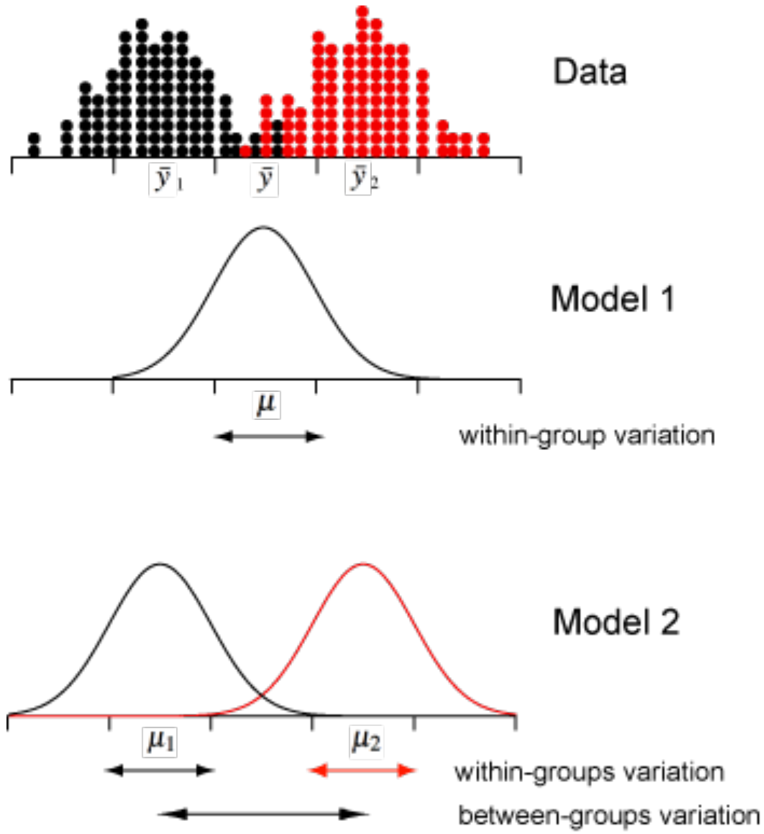
leland.wilkinson@gmail.com

Comparing

- Statistical methods exist for comparing 2 or more groups
- The classical approach is Analysis of Variance (ANOVA)
  - This method invented by Sir Ronald Fisher
  - It revolutionized industrial/scientific experiments
    - The researcher was able to examine more than one treatment at a time
  - With only two groups, results of Student's *t*-test and *F*-test are equivalent
- Multivariate Analysis of Variance (MANOVA)
  - This is ANOVA for more than one dependent variable (outcome)
- Hierarchical modeling is for nested data
  - There are several forms of this multilevel modeling

Comparing

A simple two-group comparison



Comparing

A simple two-group comparison

We compare Model 1 vs Model 2

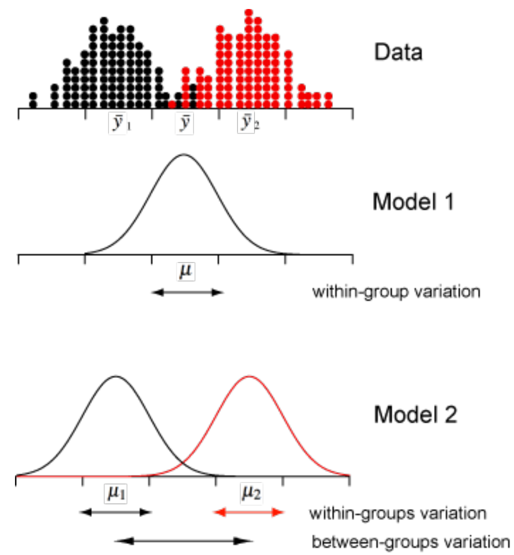
A likelihood ratio test would do for large samples

Full model (Model 2):

$$y_i = \beta_0 + \beta_1 x + \epsilon_i$$
$$= \mu + \tau + \epsilon_i$$

Restricted model (Model 1):

$$y_i = \mu + \epsilon_i$$



But for small samples, use Student's *t*-test

- Don't bother with all the unnecessarily complicated intro stat book formulas
  - They are useless
- You don't want to try this at home, folks
  - Let the stat package do it
- You want the Satterthwaite formula
  - The standard pooled formula is almost never valid on real data
  - The Satterthwaite formula gives the same answer if the variances are equal

Comparing

A simple two-group comparison

The independent groups *t*-test

Assumptions

The variable is normally distributed

The groups are independent

BUT the *t*-test is for *small n*

As David Freedman pointed out, if *n* is so small that you need a *t*-test, then the sample is too small to assess the normality assumption

And if *n* is large enough to assess normality, then you might as well use a Normal *z*-test instead of *t*

The variances are supposed to be equal.

Some say the *t*-test is robust violations of that assumption.

Then why does the Satterthwaite modification exist?

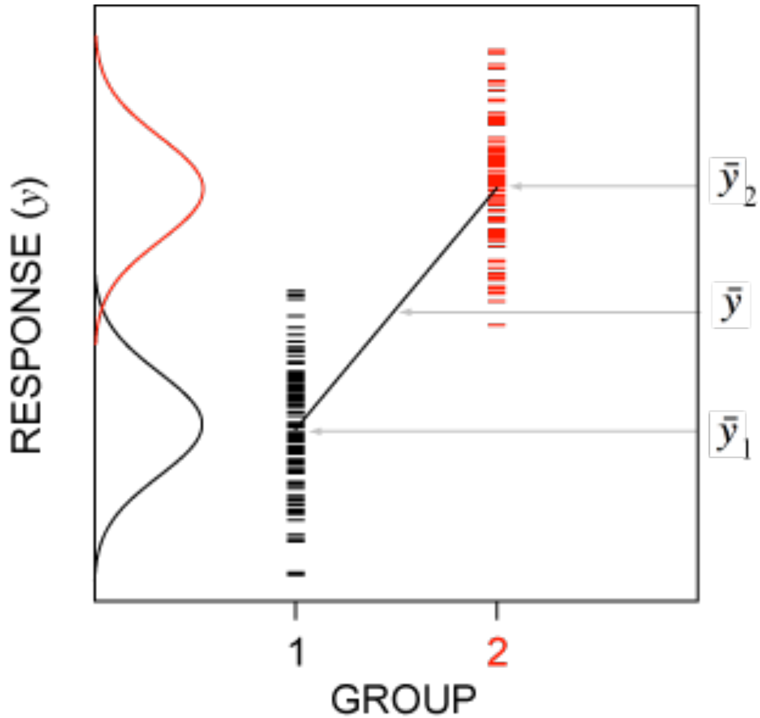
And no, the *t*-test (and *F*-test) are not robust against skewness

Comparing

Another way at looking at the independent groups test

The OLS regression model on two groups

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$



Comparing

Another way at looking at the independent groups test

Effects coding

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\mathbf{y} = \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{n_1,1} \\ y_{1,2} \\ y_{2,2} \\ \vdots \\ y_{n_2,2} \end{bmatrix} \qquad \mathbf{X} = \begin{array}{c} \begin{array}{cc} \text{grand mean} & \text{mean difference} \end{array} \\ \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{bmatrix} \end{array} \quad \mathbf{b} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} \epsilon_{1,1} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{n_1,1} \\ \epsilon_{1,2} \\ \epsilon_{2,2} \\ \vdots \\ \epsilon_{n_2,2} \end{bmatrix}$$

Comparing

Another way at looking at the independent groups test

Means coding

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\mathbf{y} = \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{n_1,1} \\ y_{1,2} \\ y_{2,2} \\ \vdots \\ y_{n_2,2} \end{bmatrix} \qquad \mathbf{X} = \begin{array}{c} \begin{array}{cc} \text{mean 1} & \text{mean 2} \end{array} \\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \end{array} \qquad \mathbf{b} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} \epsilon_{1,1} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{n_1,1} \\ \epsilon_{1,2} \\ \epsilon_{2,2} \\ \vdots \\ \epsilon_{n_2,2} \end{bmatrix}$$

Comparing

Another way at looking at the independent groups test

Hypothesis tests

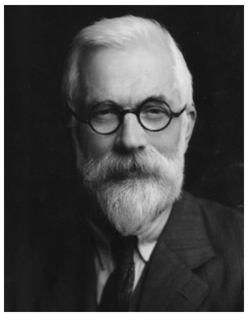
$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

Go ahead and do all the usual things

- Confidence intervals on effects coded estimates are confidence intervals on difference between cell means.
- Confidence intervals on means coded estimates are confidence intervals on cell means.
- Examine residuals
  - You want to see same variance and Normal distribution in both groups

Comparing

Analysis of Variance (ANOVA) sums of squares



$$SSB = \sum_{j=1}^{g} n_j (\hat{\mu}_j - \hat{\mu})^2$$

between groups (regression) sum of squares

$$SSW = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \hat{\mu}_j)^2$$

within groups (error) sum of squares

$$SST = \sum_{i=1}^{n} (y_i - \hat{\mu})^2$$

total sum of squares

$$MSB = SSB/(g-1)$$

mean square between groups

$$MSW = SSW/(n-g)$$

mean square within groups

$$F_{g-1,n-g} = MSB/MSW$$

F test for difference between cell means

Comparing

A simple two-group comparison

The dependent groups *t*-test

- Suppose you have repeated measures on the same subjects (e.g., pre-post)
- Then you need a *dependent t*-test
- Forget about the intro-stat textbook formulas
  - They are useless
- The same cautions apply to this situation concerning assumptions, however
- And there's a nasty gotcha
  - The dependent *t*-test takes advantage of the variance of dependent random variables

$$VAR(X + Y) = VAR(X) + VAR(Y) + 2COV(X, Y)$$

  - Actually, we're working with a difference here, so

$$VAR(X - Y) = VAR(X) + VAR(Y) - 2COV(X, Y)$$

  - So, if your measure is positively correlated across subjects, you've increased the power
  - If they are negatively correlated, however, you've decreased the power

- Ever see a researcher test whether the within-subject correlation is positive before using the dependent *t*-test?
  - I didn't think so

Comparing

A simple two-group comparison

The dependent groups *t*-test

- But it gets worse
- You don't want to use change (difference) scores for a pre-post design.
- Instead, you want an analysis of covariance with Pre as a covariate and Post as the dependent variable (more on that later)
- And if you did a Pre-Post design with Experiment and Control groups?
  - Hope you randomly assigned to treatments
  - Hope you know that the test in this case involves an interaction in a repeated measures design (we'll talk about that later)

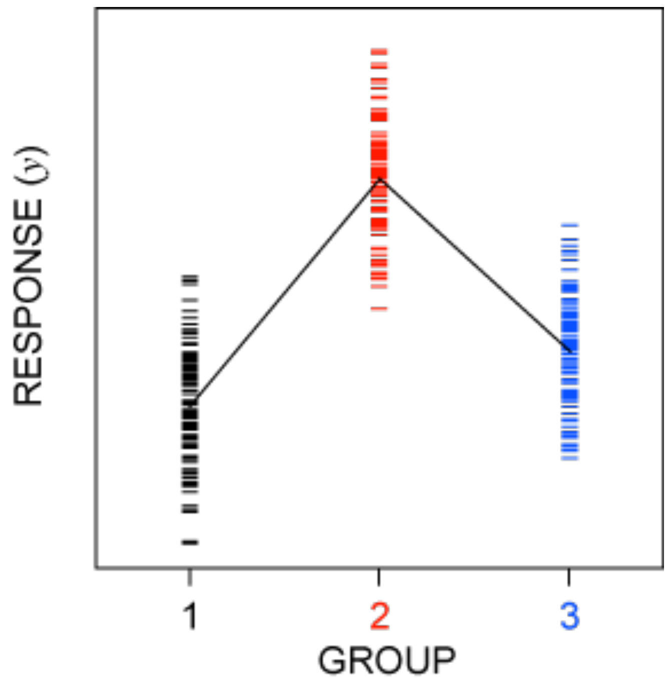
- And you thought A/B testing is simple?
  - Only market researchers and Web designers think that.

Comparing

Three groups

Same model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$



Comparing

Three groups

Means coding

$$
\mathbf{y} = \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{n_1,1} \\ y_{1,2} \\ y_{2,2} \\ \vdots \\ y_{n_2,2} \\ y_{1,3} \\ y_{2,3} \\ \vdots \\ y_{n_3,3} \end{bmatrix}
\qquad
\mathbf{X} = \begin{array}{c} \begin{array}{ccc} \text{mean 1} & \text{mean 2} & \text{mean 3} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} \end{array}
\qquad
\mathbf{b} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}
\qquad
\mathbf{e} = \begin{bmatrix} \epsilon_{1,1} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{n_1,1} \\ \epsilon_{1,2} \\ \epsilon_{2,2} \\ \vdots \\ \epsilon_{n_2,2} \\ \epsilon_{1,3} \\ \epsilon_{2,3} \\ \vdots \\ \epsilon_{n_3,3} \end{bmatrix}
$$

Comparing

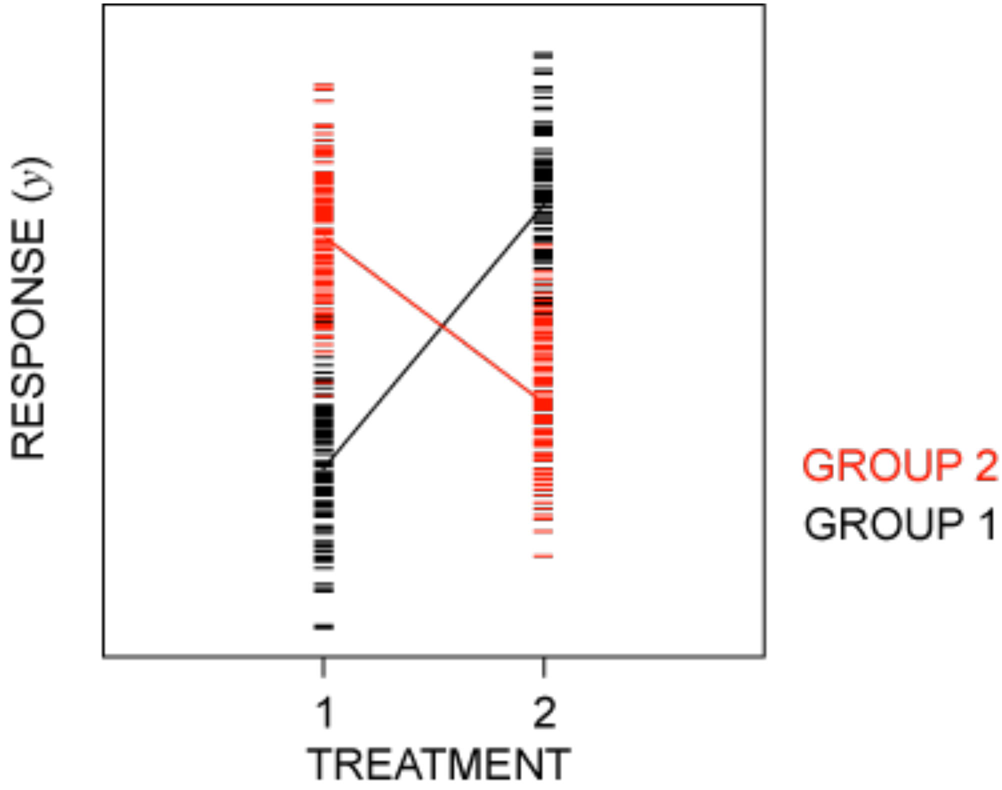
Three groups

Effects coding

$$\mathbf{y} = \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{n_1,1} \\ y_{1,2} \\ y_{2,2} \\ \vdots \\ y_{n_2,2} \\ y_{1,3} \\ y_{2,3} \\ \vdots \\ y_{n_3,3} \end{bmatrix} \quad \mathbf{X} = \begin{array}{c} \begin{array}{ccc} \text{grand mean} & \text{mean 1 vs. mean 3} & \text{mean 2 vs. mean 3} \end{array} \\ \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ \vdots & \vdots & \vdots \\ 1 & -1 & -1 \end{bmatrix} \end{array} \quad \mathbf{b} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} \epsilon_{1,1} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{n_1,1} \\ \epsilon_{1,2} \\ \epsilon_{2,2} \\ \vdots \\ \epsilon_{n_2,2} \\ \epsilon_{1,3} \\ \epsilon_{2,3} \\ \vdots \\ \epsilon_{n_3,3} \end{bmatrix}$$

Comparing

The two-way factorial model (2 x 2 design)



Comparing

The two-way factorial model (2 x 2 design)

Effects coding

$$
\mathbf{y} = \begin{bmatrix} y_{1,1,1} \\ y_{2,1,1} \\ \vdots \\ y_{n_{1,1},1,1} \\ y_{1,2,1} \\ y_{2,2,1} \\ \vdots \\ y_{n_{2,1},2,1} \\ y_{1,1,2} \\ y_{2,1,2} \\ \vdots \\ y_{n_{1,2},1,2} \\ y_{1,2,2} \\ y_{2,2,2} \\ \vdots \\ y_{n_{2,2},2,2} \end{bmatrix}
\quad
\begin{array}{c}
\begin{array}{cccc} \text{grand mean} & \text{main effect 1} & \text{main effect 2} & \text{interaction} \end{array} \\
\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & 1 \end{bmatrix}
\end{array}
\quad
\mathbf{b} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{12} \end{bmatrix}
\quad
\mathbf{e} = \begin{bmatrix} \epsilon_{1,1,1} \\ \epsilon_{2,1,1} \\ \vdots \\ \epsilon_{n_{1,1},1,1} \\ \epsilon_{1,2,1} \\ \epsilon_{2,2,1} \\ \vdots \\ \epsilon_{n_{2,1},2,1} \\ \epsilon_{1,1,2} \\ \epsilon_{2,1,2} \\ \vdots \\ \epsilon_{n_{1,2},1,2} \\ \epsilon_{1,2,2} \\ \epsilon_{2,2,1} \\ \vdots \\ \epsilon_{n_{2,2},2,2} \end{bmatrix}
$$

Comparing

Multiway factorials

- Don't even try to look at the design matrix
- Aren't you glad there's computer software for this?

Comparing

Things to consider with ANOVA

Don’t even LOOK at any lower term if it is contained in a significant interaction

FIGURE 1
Interactive Effect of Mood and Gender on Duration Estimates

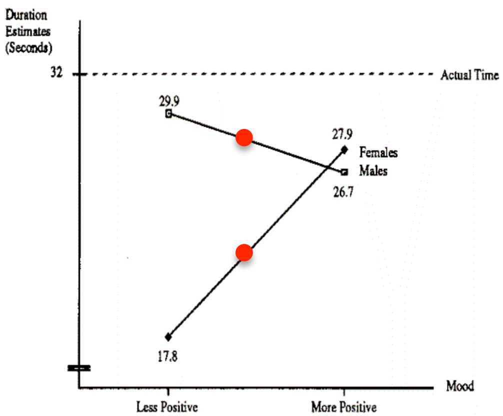


TABLE 1
Variance Analysis (ANOVA) On Duration Estimates

Source of Variation	MS	F	d.f.	p
A: Mood	35.74	.32	1,109	n.s.
B: Gender	482.76	4.29	1,109	.041
A X B Interaction	1030.18	9.16	1,109	.003
Error	112.44			

“Variance analysis found a significant main effect of gender on perceived duration (F(1,109)=4.29, p<.05).”
James J. Kellaris and Susan Powell Mantel (1994) ,"The Influence of Mood and Gender on Consumers' Time Perceptions", in NA - Advances in Consumer Research Volume 21, eds. Chris T. Allen and Deborah Roedder John, Provo, UT : Association for Consumer Research, Pages: 514-518.

BULLSHIT
The story is different for males and females

Comparing

Things to consider with ANOVA

Don't even LOOK at any lower term if it is contained in a significant interaction.

If you want to say something about main effects, you will have to do simple contrasts.

	Sum of Squares	df	Mean Square	F Ratio	Probability
Temp	3,287.111	1	3,287.111	1.208	0.283
Density	6,241.000	1	6,241.000	2.294	0.143
Salinity	51,984.722	2	25,992.361	9.554	0.001
Temp × Density	25,600.000	1	25,600.000	9.410	0.005
Temp × Salinity	368,744.056	2	184,372.028	67.773	0.000
Density × Salinity	9,852.167	2	4,926.083	1.811	0.185
Temp × Density × Salinity	54,416.167	2	27,208.083	10.001	0.001
error	65,290.667	24	2,720.444		

Comparing

Things to consider with ANOVA

- Don't trust *p* values in multiway factorials
  - Use FDR on all effects
  - Probability plot the *p* values on a uniform
- And, excuse me
  - Try explaining a 4-way interaction to someone
- The example on the right is from SYSTAT
  - I generated random data and got 2 significant effects

You can see the advantage of the ANOVA command over the MODEL statement when you have lots of factors. The equivalent MODEL statement would be as follows:

```
MODEL YIELD=CONSTANT+ A + B + C + D,
               + A*B + A*C + A*D + B*C + B*D + C*D,
               + A*B*C + A*B*D + A*C*D + B*C*D,
               + A*B*C*D
```

Here is the output:

```
DEP VAR:   YIELD     N:   32    MULTIPLE R:  .755     SQUARED MULTIPLE R:  .570
```

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
A	369800.000	1	369800.000	4.651	0.047 ★
B	1458.000	1	1458.000	0.018	0.894
C	5565.125	1	5565.125	0.070	0.795
D	172578.125	1	172578.125	2.170	0.160
A*B	87153.125	1	87153.125	1.096	0.311
A*C	137288.000	1	137288.000	1.727	0.207
A*D	328860.500	1	328860.500	4.136	0.059
B*C	61952.000	1	61952.000	0.779	0.390
B*D	3200.000	1	3200.000	0.040	0.844
C*D	3160.125	1	3160.125	0.040	0.844
A*B*C	81810.125	1	81810.125	1.029	0.326
A*B*D	4753.125	1	4753.125	0.060	0.810
A*C*D	415872.000	1	415872.000	5.230	0.036 ★
B*C*D	4.500	1	4.500	0.000	0.994
A*B*C*D	15051.125	1	15051.125	0.189	0.669
ERROR	1272247.000	16	79515.438		

We have a significant main effect for the first factor (A) plus one significant interaction (A*C*D). Let's look at the study more closely.

SYSTAT   SYSTAT - 503

Comparing

Things to consider with ANOVA

- *F* tests are generally robust against heterogeneity of variance
- But not against skewness
  - If your data are highly skewed, you are probably using wrong model
    - Counts? (you probably want Poisson)
    - Incomes? (you probably want to log the dependent variable to take care of Bill Gates)
  - Check out the next example

Comparing

Things to consider with ANOVA

Poisson ANOVA (thanks to Jerry Dallal for this final exam question)

	Sum of Squares	df	Mean Square	F Ratio	Probability
gender	18.490	1	18.490	15.448	0.000
diet	68.890	1	68.890	57.556	0.000
gender × diet	25.000	1	25.000	20.887	0.000
error	473.980	396	1.197		

ANOVA

	Coefficient	Standard Error	Lower95%	Upper95%
Constant	0.180	0.047	0.088	0.273
gender: Male	0.120	0.047	0.027	0.212
diet: W	0.315	0.047	0.222	0.407
gender: Male × diet: W	0.160	0.047	0.068	0.252

Poisson ANOVA

Comparing

Analysis of Covariance (ANCOVA)

Just throw any continuous variables you want into X

It's the same least squares model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

Here's one covariate (x) and one treatment (τ)

$$y_{ij} = \mu + \tau_j + \beta(x_{ij} - \bar{x}_j) + \epsilon_{ij}$$ (group indexed by *j*, case indexed by *i*)

We subtract the mean of the covariate ($\bar{x}_j$) out to specify deviations from cell means in the model

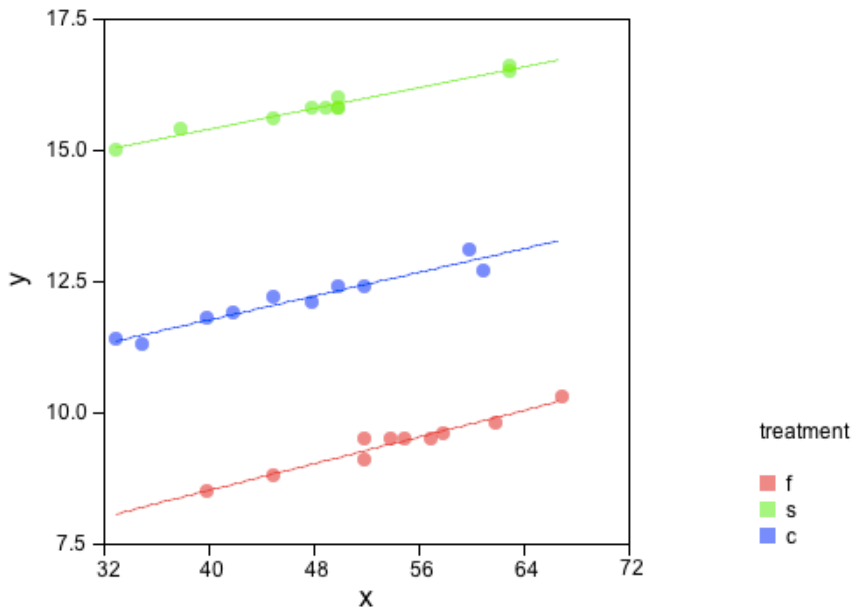
Comparing

Analysis of Covariance (ANCOVA)

Here's what we are modeling (3 groups, one covariate)

If the lines are parallel, then we can impute the effect of the treatment by looking at how vertically separated the three regression lines are

$$y_{ij} = \mu + \tau_j + \beta(x_{ij} - \bar{x}_j) + \epsilon_{ij}$$



Comparing

Things to consider with ANCOVA

- ANCOVA does not "control" for the covariate
  - it is like blocking or matching
  - regression doesn't "control" anything
  - control requires random assignment
- The separate regressions should have parallel slopes
  - if the slopes are different, add an interaction term between the covariate and the treatment
    - of course, this will make your interpretation of the results more difficult
    - this is a similar problem to testing simple effects in factorial ANOVA
  - testing this interaction term is often called testing the "parallelism assumption"
- The other usual assumptions of ANOVA still apply

Comparing

Multivariate Analysis of Variance (MANOVA)

The model is the same, except Y is now a matrix

The dimensionality of Y is *q*

$$\mathbf{Y}_{nq} = \mathbf{X}_{np}\mathbf{B}_{pq} + \mathbf{E}_{nq}$$

Estimation is the same (ordinary least squares)

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}$$

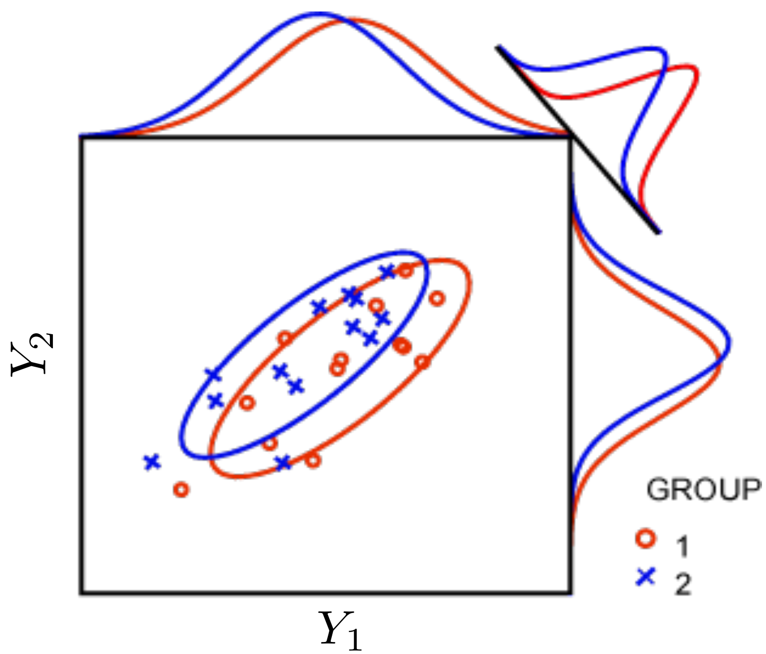
But our hypothesis tests require a multivariate distribution

We normally assume a multivariate Normal distribution

Comparing

Multivariate Analysis of Variance (MANOVA)

We seek a rotation that produces a maximum ratio of between and within groups variance



Comparing

Multivariate Analysis of Variance (MANOVA)

Testing hypotheses

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Contrast matrix

$$\mathbf{H} = \mathbf{B}'\mathbf{A}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{A}\mathbf{B}$$

Hypothesis sum of squares

$$\mathbf{G} = \mathbf{E}'\mathbf{E}$$

Error sum of squares

$$(\mathbf{H} - \lambda\mathbf{G})\mathbf{v} = \mathbf{0}$$

Characteristic equation

Comparing

Multivariate Analysis of Variance (MANOVA)

Testing hypotheses

- *Roy's Largest Root*: based the first (largest) eigenvalue
- *Wilks' Lambda*: based on the product of the reciprocal eigenvalues
- *Pillai Trace*: based on the sum of the reciprocal eigenvalues
- *Hotelling-Lawley Trace*: based on the sum of the eigenvalues

Wilks' Lambda can be transformed to exact or approximate *F*

If you don't know what an eigenvalue is, don't worry

- Most people who use statistics packages don't know either
  - But they love to use the word at cocktail parties
  - It's also called a characteristic value or latent root
  - Germans prefer the term eigenvalue
  - Malcolm Gladwell prefers the term Igon Value (Steven Pinker, NYT)

Comparing

Repeated Measures ANOVA

Use the MANOVA model (it's safer)

Testing hypotheses

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Treatments contrasts

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}$$

Measures (trials) contrasts

One can also use polynomials (linear, quadratic, cubic, …) in **C** matrix

Comparing

Repeated Measures ANOVA

- Use the MANOVA model (it's safer)
- Testing hypotheses

$$\mathbf{H} = \mathbf{C}'\mathbf{B}'\mathbf{A}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{ABC}$$ Hypothesis sum of squares

$$\mathbf{G} = \mathbf{C}'\mathbf{E}'\mathbf{EC}$$ Error sum of squares

Comparing

Repeated Measures ANOVA

Assume we have 4 groups and 3 trials.
In the one-way repeated measures model, we are interested in three tests:

- Are the 4 profiles parallel? (no group x trial interaction)
- Are all 4 profiles coincident? (no group effect)
- Are the profiles level? (no trial effect)

These tests are done in sequence.

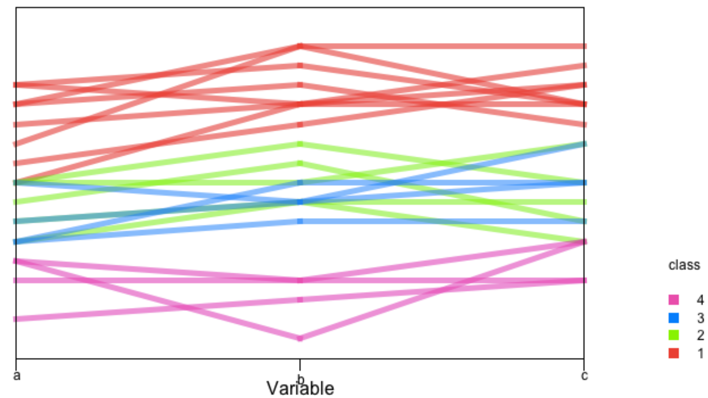
1) If the 4 profiles are parallel, then we can go on to compare means across profiles to see if they are coincident. Otherwise, there is an interaction between the trials factor and the grouping factor and we have to stop there.

2) If the 4 profiles are coincident, then we can go on to test whether they are level. If not, then there is a groups effect and we have to stop there.

3) If they are level, then there is no trials effect.

Comparing

Repeated Measures ANOVA



Effect		F Ratio	df1	df2	Probability
Within Subjects	Constant (multivariate)	6.191	2	16	0.010
	Linear	12.226	1	17	0.003
	Quadratic	0.310	1	17	0.585
	class (multivariate)	1.773	6	32	0.136
	Linear	1.010	3	17	0.413
	Quadratic	2.818	3	17	0.070
Between Subjects	Constant	3,456.884	1	17	0.000
	class	70.927	3	17	0.000