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

Data

Model 1

Model 2

within-group variation

within-groups variation

between-groups variation

Copyright © 2016 Leland Wilkinson
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):
\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]
\[ = \mu + \tau + \epsilon_i \]

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

\[ y = Xb + e \]
Comparing

Another way at looking at the independent groups test

Effects coding

\[ y = Xb + e \]

\[ y = \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{n,1} \\ y_{1,2} \\ y_{2,2} \\ \vdots \\ y_{n,2} \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{bmatrix} \quad b = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad e = \begin{bmatrix} \epsilon_{1,1} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{n,1} \\ \epsilon_{1,2} \\ \epsilon_{2,2} \\ \vdots \\ \epsilon_{n,2} \end{bmatrix} \]
Comparing

Another way at looking at the independent groups test

Means coding

\[ y = Xb + e \]

\[
\begin{bmatrix}
  y_{1,1} \\
  y_{2,1} \\
  \vdots \\
  y_{n,1,1} \\
  y_{1,2} \\
  y_{2,2} \\
  \vdots \\
  y_{n,2,2}
\end{bmatrix} =
\begin{bmatrix}
  1 & 0 \\
  1 & 0 \\
  \vdots & \vdots \\
  1 & 0 \\
  0 & 1 \\
  0 & 1 \\
  \vdots & \vdots \\
  0 & 1
\end{bmatrix}
\begin{bmatrix}
  \mu_1 \\
  \mu_2
\end{bmatrix} +
\begin{bmatrix}
  \epsilon_{1,1} \\
  \epsilon_{2,1} \\
  \vdots \\
  \epsilon_{n,1,1} \\
  \epsilon_{1,2} \\
  \epsilon_{2,2} \\
  \vdots \\
  \epsilon_{n,2,2}
\end{bmatrix}
\]
Comparing

Another way at looking at the independent groups test

Hypothesis tests

\[ y = Xb + e \]

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
\[
\text{VAR}(X + Y) = \text{VAR}(X) + \text{VAR}(Y) + 2\text{COV}(X, Y)
\]
Actually, we’re working with a difference here, so
\[
\text{VAR}(X - Y) = \text{VAR}(X) + \text{VAR}(Y) - 2\text{COV}(X, Y)
\]
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

\[ y = Xb + e \]
Comparing

Three groups

Means coding

\[ y = \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{n_1,1} \\ y_{1,2} \\ y_{2,2} \\ \vdots \\ y_{n_2,2} \\ y_{1,3} \\ y_{2,3} \\ \vdots \\ y_{n_3,3} \end{bmatrix} \]

\[ X = \begin{bmatrix} 1 & 0 & 0 & & & & & & \end{bmatrix} \]

\[ b = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \]

\[ e = \begin{bmatrix} \epsilon_{1,1} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{n_1,1} \\ \epsilon_{1,2} \\ \epsilon_{2,2} \\ \vdots \\ \epsilon_{n_2,2} \\ \epsilon_{1,3} \\ \epsilon_{2,3} \\ \vdots \\ \epsilon_{n_3,3} \end{bmatrix} \]
Comparing

Three groups

Effects coding

\[
\begin{align*}
\mathbf{y} &= \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{n_1,1} \\ y_{1,2} \\ y_{2,2} \\ \vdots \\ y_{n_2,2} \\ y_{1,3} \\ y_{2,3} \\ \vdots \\ y_{n_3,3} \end{bmatrix} & & \mathbf{X} &= \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ \vdots & \vdots & \vdots \\ 1 & -1 & -1 \end{bmatrix} & & \mathbf{b} &= \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} & & \mathbf{e} &= \begin{bmatrix} \epsilon_{1,1} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{n_1,1} \\ \epsilon_{1,2} \\ \epsilon_{2,2} \\ \vdots \\ \epsilon_{n_2,2} \\ \epsilon_{1,3} \\ \epsilon_{2,3} \\ \vdots \\ \epsilon_{n_3,3} \end{bmatrix}
\end{align*}
\]
Comparing

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

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

Effects coding

\[
\begin{align*}
\mathbf{y} &= \begin{bmatrix}
y_{1,1,1} \\
y_{2,1,1} \\
\vdots \\
y_{n_{1,1,1,1}} \\
y_{1,2,1} \\
y_{2,2,1} \\
\vdots \\
y_{n_{2,1,2,1}} \\
y_{1,1,2} \\
y_{2,1,2} \\
\vdots \\
y_{n_{1,2,1,2}} \\
y_{1,2,2} \\
y_{2,2,2} \\
\vdots \\
y_{n_{2,2,2,2}}
\end{bmatrix} \\
\mathbf{X} &= \begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
\vdots & \vdots & \vdots & \vdots \\
1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 \\
1 & -1 & 1 & -1 \\
\vdots & \vdots & \vdots & \vdots \\
1 & 1 & -1 & -1 \\
1 & 1 & -1 & -1 \\
\vdots & \vdots & \vdots & \vdots \\
1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 \\
1 & -1 & -1 & 1 \\
\vdots & \vdots & \vdots & \vdots \\
1 & -1 & -1 & 1
\end{bmatrix} \\
\mathbf{b} &= \begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_{12}
\end{bmatrix} \\
\mathbf{e} &= \begin{bmatrix}
\epsilon_{1,1,1,1} \\
\epsilon_{2,1,1,1} \\
\vdots \\
\epsilon_{n_{1,1,1,1},1,1} \\
\epsilon_{1,2,1,1} \\
\epsilon_{2,2,1,1} \\
\vdots \\
\epsilon_{n_{2,1,2,1},1,2,1} \\
\epsilon_{1,1,2,1,2} \\
\epsilon_{2,1,2,1,2} \\
\vdots \\
\epsilon_{n_{1,2,1,2},1,2,1} \\
\epsilon_{1,2,2,1,2} \\
\epsilon_{2,2,2,1,2} \\
\vdots \\
\epsilon_{n_{2,2,2,2},1,2,1}
\end{bmatrix}
\end{align*}
\]
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

“Variance analysis found a significant main effect of gender on perceived duration ($F(1,109)=4.29$, $p<.05$).”

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.

<table>
<thead>
<tr>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp</td>
<td>1</td>
<td>3,287.111</td>
<td>1.208</td>
<td>0.283</td>
</tr>
<tr>
<td>Density</td>
<td>1</td>
<td>6,241.000</td>
<td>2.294</td>
<td>0.143</td>
</tr>
<tr>
<td>Salinity</td>
<td>2</td>
<td>25,992.361</td>
<td>9.554</td>
<td>0.001</td>
</tr>
<tr>
<td>Temp × Density</td>
<td>1</td>
<td>25,600.000</td>
<td>9.410</td>
<td>0.005</td>
</tr>
<tr>
<td>Temp × Salinity</td>
<td>2</td>
<td>184,372.028</td>
<td>67.773</td>
<td>0.000</td>
</tr>
<tr>
<td>Density × Salinity</td>
<td>2</td>
<td>4,926.083</td>
<td>1.811</td>
<td>0.185</td>
</tr>
<tr>
<td>Temp × Density × Salinity</td>
<td>2</td>
<td>27,208.083</td>
<td>10.001</td>
<td>0.001</td>
</tr>
<tr>
<td>error</td>
<td>24</td>
<td>2,720.444</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
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)

<table>
<thead>
<tr>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>18.490</td>
<td>18.490</td>
<td>15.448</td>
<td>0.000</td>
</tr>
<tr>
<td>diet</td>
<td>68.890</td>
<td>68.890</td>
<td>57.556</td>
<td>0.000</td>
</tr>
<tr>
<td>gender × diet</td>
<td>25.000</td>
<td>25.000</td>
<td>20.887</td>
<td>0.000</td>
</tr>
<tr>
<td>error</td>
<td>473.980</td>
<td>396</td>
<td>1.197</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Lower95%</th>
<th>Upper95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.180</td>
<td>0.047</td>
<td>0.088</td>
</tr>
<tr>
<td>gender: Male</td>
<td>0.120</td>
<td>0.047</td>
<td>0.027</td>
</tr>
<tr>
<td>diet: W</td>
<td>0.315</td>
<td>0.047</td>
<td>0.222</td>
</tr>
<tr>
<td>gender: Male × diet: W</td>
<td>0.160</td>
<td>0.047</td>
<td>0.068</td>
</tr>
</tbody>
</table>
Comparing

Analysis of Covariance (ANCOVA)

Just throw any continuous variables you want into \( X \)

It’s the same least squares model

\[
y = Xb + e
\]

Here’s one covariate \((x)\) and one treatment \((\tau)\)

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

$$Y_{nq} = X_{np}B_{pq} + E_{nq}$$

Estimation is the same (ordinary least squares)

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

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

Copyright © 2016 Leland Wilkinson
Comparing

Multivariate Analysis of Variance (MANOVA)

Testing hypotheses

\[ A = \begin{bmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1
\end{bmatrix} \]

Contrast matrix

\[ H = B' A' (X' X)^{-1} A B \]

Hypothesis sum of squares

\[ G = E' E \]

Error sum of squares

\[ (H - \lambda G)v = 0 \]

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

\[
A = \begin{bmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 1
\end{bmatrix}
\]

Treatments contrasts

\[
C = \begin{bmatrix}
1 & -1 & 0 & \cdots & 0 & 0 \\
0 & 1 & -1 & \cdots & 0 & 0 \\
0 & 0 & 0 & \cdots & 1 & -1
\end{bmatrix}
\]

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

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

\[ G = C'E'EC \]  
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

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