Modern alternatives to ANOVA

Jonty Rougier
School of Mathematics

j.c.rougier@bristol.ac.uk

Population Health Sciences meeting, Apr 2018

ANalysis Of VAriance

For a *one-way layout*. Your data look like this (in R):

Calorie	Variety
175	Beef
173	Meat
144	Poultry
132	Beef
94	Poultry
149	Beef
179	Meat
180	Specialty
102	Poultry
135	Poultry
138	Meat

etc; where `Calorie` is a measurement, `Variety` is a factor (type of hotdog in this case), and `Beef`, `Meat` etc are levels of the factor.

When you should use 'classical' ANOVA

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

where μ_i is the population expectation for the ith level, AND

1. The underlying distributions of the measurements are Normal (Gaussian) for each level, AND
2. The variances of these distributions are all the same, AND
3. The alternative hypothesis is

$$H_1 : \exists i, j : \mu_i \neq \mu_j.$$

When you should use 'classical' ANOVA

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

where μ_i is the population expectation for the ith level, AND

1. The underlying distributions of the measurements are Normal (Gaussian) for each level, AND
2. The variances of these distributions are all the same, AND
3. The alternative hypothesis is

$$H_1 : \exists i, j : \mu_i \neq \mu_j.$$

Otherwise, you will have to use something else.

- 1 or 2 don't hold ▸ Example If (3) holds, Kruskal-Wallis test; otherwise several two-sample Wilcoxon rank sum tests.

When you should use 'classical' ANOVA

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

where μ_i is the population expectation for the ith level, AND

1. The underlying distributions of the measurements are Normal (Gaussian) for each level, AND
2. The variances of these distributions are all the same, AND
3. The alternative hypothesis is

$$H_1 : \exists i, j : \mu_i \neq \mu_j.$$

Otherwise, you will have to use something else.

- 1 or 2 don't hold ▸ Example If (3) holds, Kruskal-Wallis test; otherwise several two-sample Wilcoxon rank sum tests.
- 3 doesn't hold. If (1) holds, several two-sample t-tests (with unequal variances if (2) doesn't hold); otherwise several two-sample Wilcoxon rank sum tests.

Several two-sample tests

Instead of

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{versus} \quad \exists i, j : \mu_i \neq \mu_j,$$

which is just 1 test, do

$$\begin{aligned}
H_0^{12} &: \mu_1 = \mu_2 \quad \text{versus} \quad \mu_1 \neq \mu_2 \\
H_0^{13} &: \mu_1 = \mu_3 \quad \text{versus} \quad \mu_1 \neq \mu_3 \\
&\vdots \\
H_0^{1k} &: \mu_1 = \mu_k \quad \text{versus} \quad \mu_1 \neq \mu_k \\
H_0^{23} &: \mu_2 = \mu_3 \quad \text{versus} \quad \mu_2 \neq \mu_3 \\
&\vdots \\
H_0^{k-1,k} &: \mu_{k-1} = \mu_k \quad \text{versus} \quad \mu_{k-1} \neq \mu_k
\end{aligned}$$

which is $k(k-1)/2$ tests.

Several two-sample tests

Instead of

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{versus} \quad \exists i, j : \mu_i \neq \mu_j,$$

which is just 1 test, do

$$\begin{aligned}
H_0^{12} &: \mu_1 = \mu_2 \quad \text{versus} \quad \mu_1 \neq \mu_2 \\
H_0^{13} &: \mu_1 = \mu_3 \quad \text{versus} \quad \mu_1 \neq \mu_3 \\
&\vdots \\
H_0^{1k} &: \mu_1 = \mu_k \quad \text{versus} \quad \mu_1 \neq \mu_k \\
H_0^{23} &: \mu_2 = \mu_3 \quad \text{versus} \quad \mu_2 \neq \mu_3 \\
&\vdots \\
H_0^{k-1,k} &: \mu_{k-1} = \mu_k \quad \text{versus} \quad \mu_{k-1} \neq \mu_k
\end{aligned}$$

which is $k(k-1)/2$ tests. The hypotheses you reject are interesting. How to control the error rate when *multiple testing?*

Error rates: all done with p-values

- With a **single hypothesis test**, the Type 1 error α is the probability of rejecting H_0 when it is true.

  If $p \leq \alpha$ then H_0 is *rejected at a significance level of* α.

Error rates: all done with p-values

- With a **single hypothesis test**, the Type 1 error α is the probability of rejecting H_0 when it is true.

  If $p \leq \alpha$ then H_0 is *rejected at a significance level of* α.

- With **multiple hypothesis tests**, the Family-wise Error Rate (FWER) is the probability of rejecting at least one H_0 when it is true.

  If $p_i^* \leq \alpha$ then H_0^i is *rejected at a FWER of* α.

Error rates: all done with p-values

- With a **single hypothesis test**, the Type 1 error α is the probability of rejecting H_0 when it is true.

  If $p \leq \alpha$ then H_0 is *rejected at a significance level of* α.

- With **multiple hypothesis tests**, the Family-wise Error Rate (FWER) is the probability of rejecting at least one H_0 when it is true.

  If $p_i^* \leq \alpha$ then H_0^i is *rejected at a FWER of* α.

- A transformation is applies to the individual p-values $p_1, \ldots, p_m$ to derive $p_1^*, \ldots, p_m^*$. This is based on the Holm procedure.

Error rates: all done with p-values

- With a **single hypothesis test**, the Type 1 error α is the probability of rejecting H_0 when it is true.

  If $p \leq \alpha$ then H_0 is *rejected at a significance level of* α.

- With **multiple hypothesis tests**, the Family-wise Error Rate (FWER) is the probability of rejecting at least one H_0 when it is true.

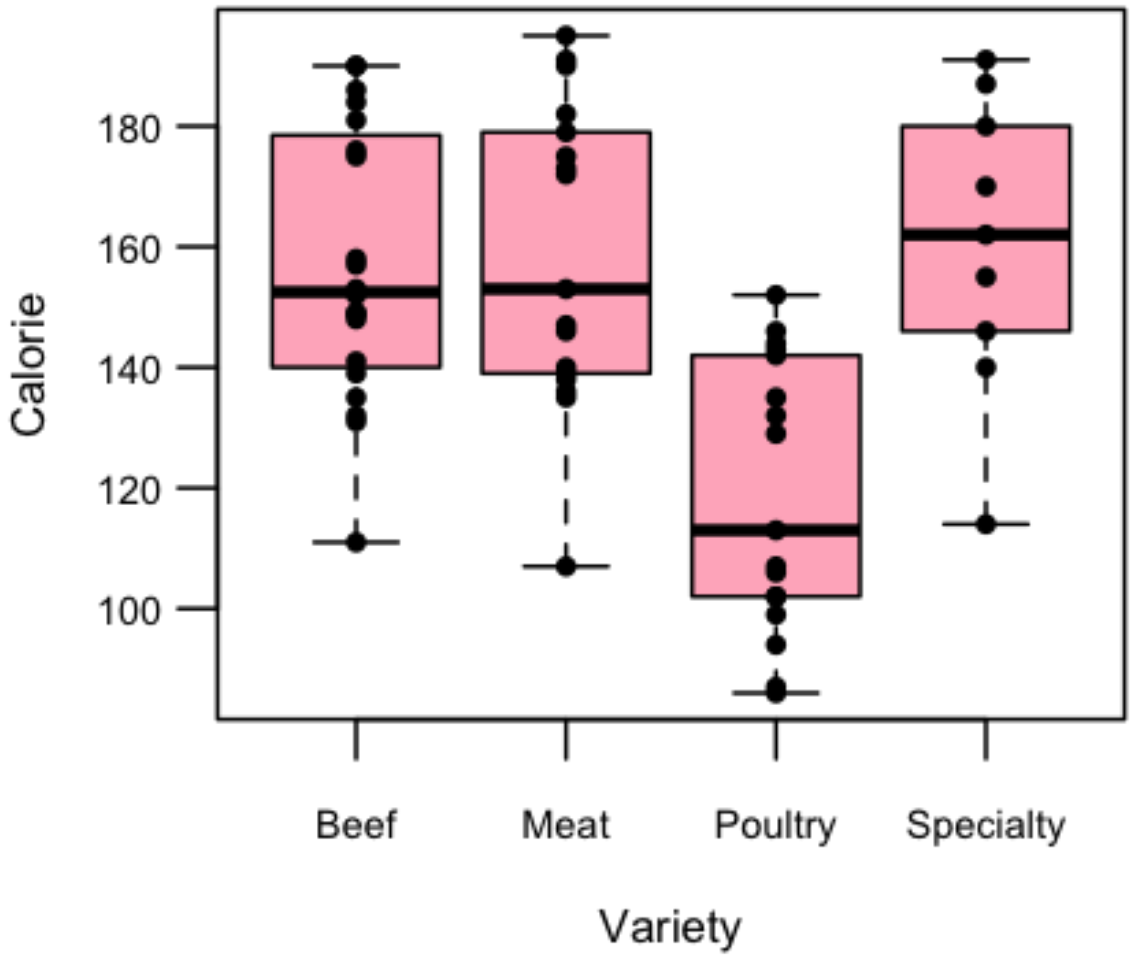
  If $p_i^* \leq \alpha$ then H_0^i is *rejected at a FWER of* α.

- A transformation is applies to the individual p-values $p_1, \ldots, p_m$ to derive $p_1^*, \ldots, p_m^*$. This is based on the Holm procedure.

In practice: compute $p_1, \ldots, p_m$; transform to $p_1^*, \ldots, p_m^*$; and highlight those i for which $p_i^* \leq 0.05$ (or some other conventional significance level).

Example: calories in hotdogs

Here are the measurements, by level (*always* draw this picture):



Example: calories in hotdogs (cont)

Suppose that we are interested in identifying interesting pairwise differences. We say, "It is clear, from inspecting the measurements, that `Beef`, `Meat`, and `Specialty` are similar, but `Poultry` is different." A ~~boneheaded~~ reviewer asks for p-values to back this up.

Example: calories in hotdogs (cont)

Suppose that we are interested in identifying interesting pairwise differences. We say, "It is clear, from inspecting the measurements, that `Beef`, `Meat`, and `Specialty` are similar, but `Poultry` is different." A ~~boneheaded~~ reviewer asks for p-values to back this up.

- Conditions (1) and (2) both appear to hold, so generate $m = 6$ p-values using two-sample t-tests with equal variances.

Example: calories in hotdogs (cont)

Suppose that we are interested in identifying interesting pairwise differences. We say, "It is clear, from inspecting the measurements, that `Beef`, `Meat`, and `Specialty` are similar, but `Poultry` is different." A ~~boneheaded~~ reviewer asks for p-values to back this up.

- Conditions (1) and (2) both appear to hold, so generate $m = 6$ p-values using two-sample t-tests with equal variances.
- The original and transformed values can be displayed in a square array, missing its diagonal:

	Beef	Meat	Poult.	Spec.
Beef		1.00000	0.00007	1.00000
Meat	0.81499		0.00015	1.00000
Poultry	0.00001	0.00003		0.00088
Specialty	0.69516	0.85961	0.00022	

with the original p-values in the lower-left, and the transformed ones in the top-right ($p_i \leq p_i^*$).

Example: calories in hotdogs (cont)

Suppose that we are interested in identifying interesting pairwise differences. We say, "It is clear, from inspecting the measurements, that `Beef`, `Meat`, and `Specialty` are similar, but `Poultry` is different." A ~~boneheaded~~ reviewer asks for p-values to back this up.

- The original and transformed values can be displayed in a square array, missing its diagonal:

	Beef	Meat	Poult.	Spec.
Beef		1.00000	0.00007	1.00000
Meat	0.81499		0.00015	1.00000
Poultry	0.00001	0.00003		0.00088
Specialty	0.69516	0.85961	0.00022	

  with the original p-values in the lower-left, and the transformed ones in the top-right ($p_i \leq p_i^*$).
- *We reject hypotheses* H_0^{13}, H_0^{23}, *and* H_0^{34} *at a FWER of* 5% *(see Table for details).*

Conservative procedures

- The Holm procedure will control the FWER with no additional conditions, but it is conservative. Crudely, it can raise p-values too much, leading to fewer rejections.
- Other procedures for controlling the FWER are less conservative, but only valid under additional conditions.

Conservative procedures

- The Holm procedure will control the FWER with no additional conditions, but it is conservative. Crudely, it can raise p-values too much, leading to fewer rejections.
- Other procedures for controlling the FWER are less conservative, but only valid under additional conditions.
- If you have a lot of levels (i.e. the number of pairwise comparisons is large), then Holm and other procedures can be *very* conservative.
- In this case (and possibly for other reasons) you may want to switch to controlling the False Discovery Rate (FDR). This is common in -omics.

  A theoretical result states that FDR $\leq$ FWER, and so there will typically be more rejections for the same threshold.

References

Lehmann and Romano (2005, ch. 9) has an introduction to the theory of multiple testing. Holm (1979) has the Holm procedure. Wright (1992) has the equations for adjusting p-values for the FWER, FDR, and a few others. If you want to control the FDR, then Benjamini and Hochberg (1995) is the original article.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Lehmann, E. and Romano, J. (2005). *Testing Statistical Hypotheses*. New York: Springer, 3rd edition.

Wright, P. (1992). Adjusted *P*-values for simultaneous inference. *Biometrics*, 48(4):1005–1013.

Power: the elephant in the room

It is appropriate to end on a note of caution.

- For any statistical model, there are an uncountable number of possible significance procedures, each one delivering a different p-value. These occupy a spectrum

$$\text{useless} \longrightarrow \text{powerful}.$$

- A 'naked' p-value conveys nothing about where on this spectrum it lies. To address this requires the calculation of power with respect to an alternative hypothesis, similar to the Neyman-Pearson approach to hypothesis testing.

Power: the elephant in the room

It is appropriate to end on a note of caution.

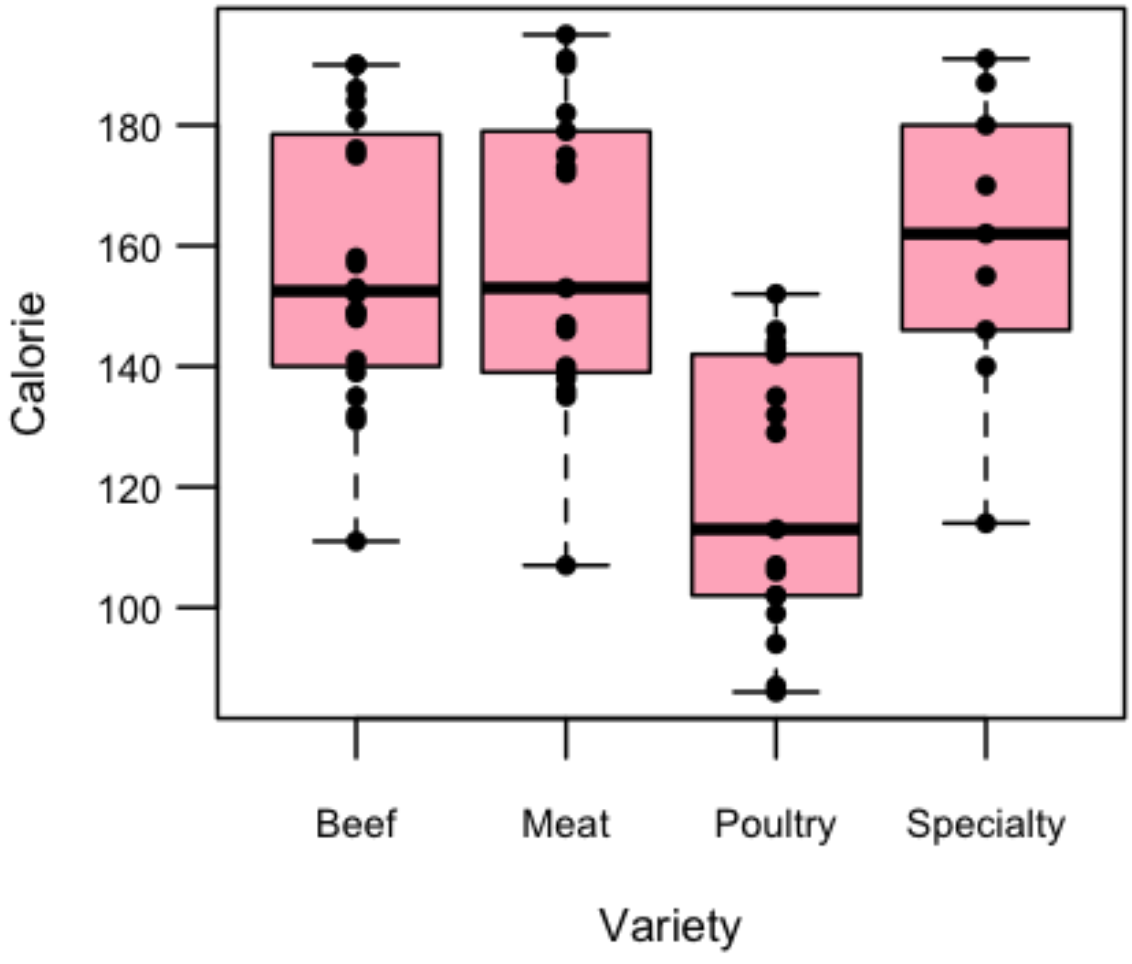
- For any statistical model, there are an uncountable number of possible significance procedures, each one delivering a different p-value. These occupy a spectrum

  $$\text{useless} \longrightarrow \text{powerful}.$$

- A 'naked' p-value conveys nothing about where on this spectrum it lies. To address this requires the calculation of power with respect to an alternative hypothesis, similar to the Neyman-Pearson approach to hypothesis testing.
- Fully parametic models, e.g. those that can be analysed using classical ANOVA or two-sample t-tests, can be evaluated for power. But non-parametric models evaluated by permutation tests, e.g. the Wilcoxon rank sum test, cannot.
- If you think it is OK to ignore issues of power when producing p-vaues, then you might like to reflect that much of the current 'crisis of reproducibility' in statistical science is due to ignorance and under-powered tests. **Don't be part of the problem!**

Failure of the Normal conditions

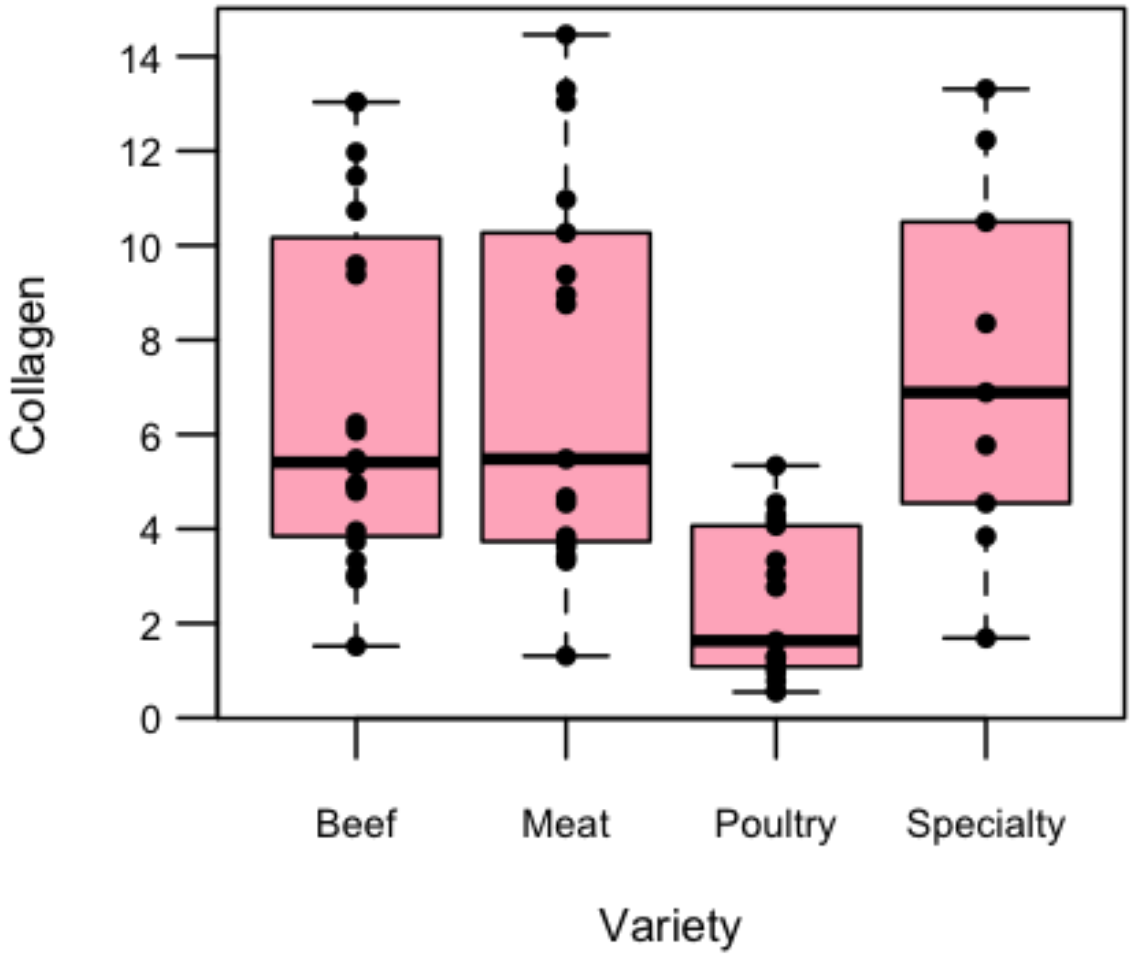
Here's one where the normal conditions appear to hold:



◂ Back again

Failure of the Normal conditions

And here's one where they don't hold:



◂ Back again