Analysis of Variance

Bret Hanlon and Bret Larget

Department of Statistics
University of Wisconsin—Madison

November 22–November 29, 2011

Cuckoo Birds

Case Study

- Cuckoo birds have a behavior in which they lay their eggs in other birds nests.
- The other birds then raise and care for the newly hatched cuckoos.
- Cuckoos return year after year to the same territory and lay their eggs in the nests of a particular host species.
- Furthermore, cuckoos appear to mate only within their territory.
- Therefore, geographical sub-species are developed, each with a dominant foster-parent species.
- A general question is, are the eggs of the different sub-species adapted to a particular foster-parent species?
- Specifically, we can ask, are the mean lengths of the cuckoo eggs the same in the different sub-species?

Cuckoo Bird Egg Length Distribution
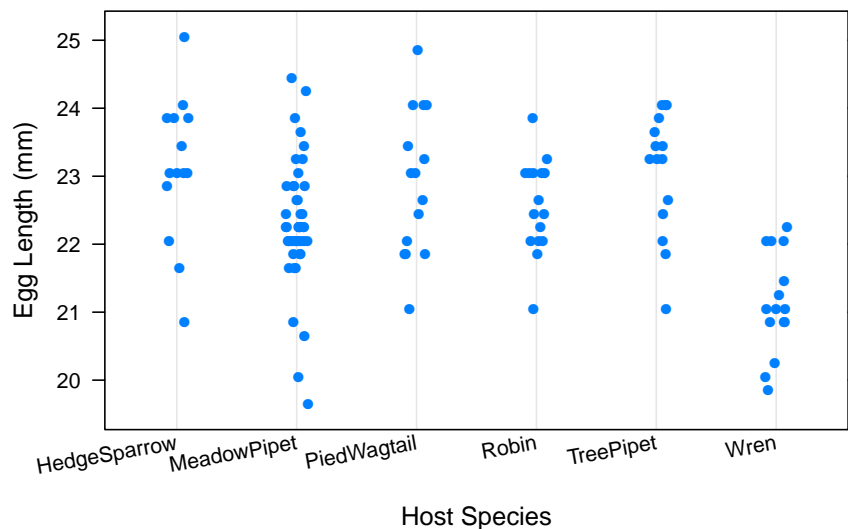
Comparing More than Two Populations

- We have developed both t and nonparametric methods for inference for comparing means from two populations.
- What if there are three or more populations?
- It is not valid to simply make all possible pairwise comparisons:
- with three populations, there are three such comparisons, with four there are six, and the number increases rapidly.
- The comparisons are not all independent: the data used to estimate the differences between the pair of populations 1 and 2 and the pair of populations 1 and 3 use the same sample from population 1.
- When estimating differences with confidence, we may be concerned about the confidence we ought to have that all differences are in their respective intervals.
- For testing, there are many simultaneous tests to consider.

What to do?

Hypotheses

- The common approach to this problem is based on a single null hypothesis

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

  versus the alternative hypothesis that the means are not all the same (so that there are at least two means that differ) where there are k groups.

- If there is evidence against the null hypothesis, then further inference is carried out to examine specific comparisons of interest.

Illustrative Example

- The dot plots show two cases of three samples, each of size five.
- The sample means are respectively 180, 220, and 200 in both cases.
- The left plot appears to show differences in the mean; evidence for this in the right plot appears weaker.
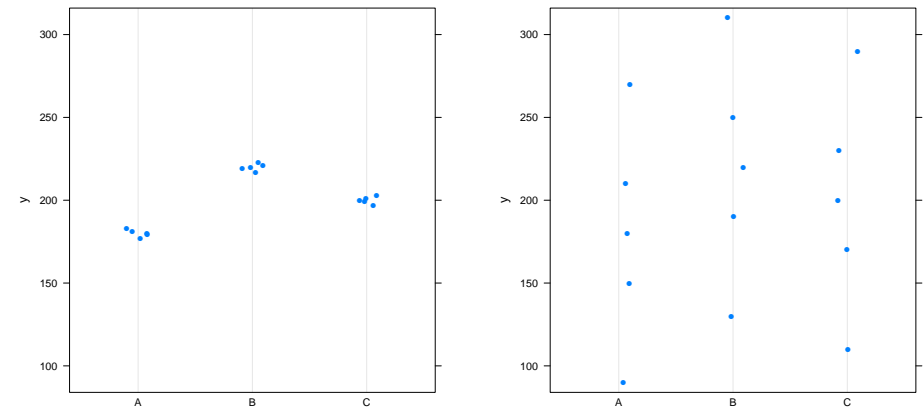
Analysis of Variance

- The previous example suggests an approach that involves comparing variances;
- If *variation among sample means is large relative to variation within samples*, then there is evidence against $H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_k$.
- If *variation among sample means is small relative to variation within samples*, then the data is consistent with $H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_k$.
- The approach of testing $H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_k$ on the basis of comparing variation among and within samples is called Analysis of Variance, or ANOVA.

ANOVA Table Concept

- To test the previous hypothesis, we construct a test statistic that is a ratio of two different and independent estimates of an assumed common variance among populations, σ^2.
- The numerator estimate is based on sample means and variation *among groups*.
- The denominator estimate is based on variation *within samples*.
- If the null hypothesis is true, then we expect this ratio to be close to one (but with random sampling, it may be somewhat greater).
- If the null hypothesis is false, then the estimate in the numerator is likely to be much larger than the estimate in the denominator, and the test statistic may be much larger than one than can be explained by chance variation alone.
- An ANOVA Table is simply an accounting method for calculating a complicated test statistic.
- The following several slides develop the notation underlying this theory.

Notation

- There are k populations.
- The ith observation is Y_i which is in the $j(i)$th sample.
- We write $j(i)$ to indicate the group associated with observation i.
- We let i vary from 1 to n, the total number of observations.
- j varies from 1 to k, the total number of populations/samples.
- There are a total of n observations with n_j observations in sample j.

$$n = n_1 + \cdots + n_k$$

Sample Mean and SD

- The sample mean for a group is the sum of all observations in the group divided by the number in the group.
- The notation $i : j(i) = j$ means *all i such that observation i is in group j*.
- The sample mean in group j is:

$$\bar{Y}_j = \frac{\sum\limits_{i:j(i)=j} Y_i}{n_j}$$

and the sample standard deviation in group j is

$$s_j = \sqrt{\frac{\sum\limits_{i:j(i)=j} (Y_i - \bar{Y}_j)^2}{n_j - 1}}$$

Grand Mean

- The *grand mean* $\bar{Y}$ is the mean of all observations.
- Note that the grand mean

$$\bar{Y} = \sum_{j=1}^{k} \left(\frac{n_j}{n} \right) \bar{Y}_j$$

is *the weighted average of the sample means, weighted by sample size.*

Modeling Assumptions

We make the following modeling assumptions:

- All observations Y_i are independent.
- $E(Y_i) = \mu_{j(i)}$, where $\mu_{j(i)}$ is the mean of population j from which observation i was drawn.
- $\text{Var}(Y_i) = \sigma^2_{j(i)}$, where $\sigma^2_{j(i)}$ is the variance of population j.

We will also often make the following two additional assumptions:

- all population variances are equal: $\sigma_j^2 = \sigma^2$ for all j;
- all observations are normally distributed: $Y_i \sim \text{N}(\mu_{j(i)}, \sigma_{j(i)})$

Distributions of the Sample Means

With the first set of assumptions, note that

- $E(\bar{Y}_j) = \mu_j$ and $Var(\bar{Y}_j) = \frac{\sigma_j^2}{n_j}$

and additionally, if the second set of assumptions are made, then

- $\bar{Y}_j \sim N\left(\mu_j, \frac{\sigma}{\sqrt{n_j}}\right)$

Variation Among Samples

- We use this formula for the variation among sample means:

$$\sum_{j=1}^{k} n_j(\bar{Y}_j - \bar{Y})^2$$

which is *a weighted sum of squared deviations of sample means from the grand mean, weighted by sample size*.

- Under the assumptions of independence and equal variances,

$$E\left(\sum_{j=1}^{k} n_j(\bar{Y}_j - \bar{Y})^2\right) = (k-1)\sigma^2 + \sum_{j=1}^{k} n_j(\mu_j - \mu)^2$$

where

$$\mu = \frac{\sum_{j=1}^{k} n_j\mu_j}{n}$$

is the expected value of the grand mean $\bar{Y}$.

Variation Among Samples (cont.)

- The sum

$$\sum_{j=1}^{k} n_j(\bar{Y}_j - \bar{Y})^2$$

is called the *group sum of squares*.
- If the null hypothesis $H_0\colon \mu_1 = \mu_2 = \cdots = \mu_k$ is true, then $\sum_{j=1}^{k} n_j(\mu_j - \mu)^2 = 0$ and

$$E\left(\sum_{j=1}^{k} n_j(\bar{Y}_j - \bar{Y})^2\right) = (k-1)\sigma^2$$

- This suggests defining

$$MS_{\text{groups}} = \frac{\sum_{j=1}^{k} n_j(\bar{Y}_j - \bar{Y})^2}{k-1}$$

to be the *group mean square*.
- If the null hypothesis is true, then $E(MS_{\text{groups}}) = \sigma^2$; otherwise, $E(MS_{\text{groups}}) = \sigma^2 + \sum_{j=1}^{k} n_j(\mu_j - \mu)^2/(k-1) > \sigma^2$.

Variation Within Samples

- For each sample, the sample variance

$$s_j^2 = \frac{\sum_{i:j(i)=j}(Y_i - \bar{Y}_j)^2}{n_j - 1}$$

is an estimate of that population's variance, σ_j^2.
- Under the assumptions of equal variance and independence, each s_j^2 is then an independent estimate of σ^2.
- The formula

$$\sum_{j=1}^{k} (n_j - 1)s_j^2$$

is the sum of all squared deviations from individual sample means and has expected value

$$E\left(\sum_{j=1}^{k} (n_j - 1)s_j^2\right) = (n-k)\sigma^2$$

Variation Within Samples (cont.)

- The mean square error formula

$$\mathrm{MS}_{\mathrm{error}} = \frac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{n - k}$$

  is *a weighted average of the sample variances, weighted by degrees of freedom*.

- Notice that $\mathrm{E}(\mathrm{MS}_{\mathrm{error}}) = \sigma^2$ always: it is true when $H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_k$ is true, but also when H_0 is false.

The F Test Statistic

- We have developed two separate formulas for variation among and within samples, each based on a different *mean square*:
  - $\mathrm{MS}_{\mathrm{groups}}$ measures variation among groups;
  - $\mathrm{MS}_{\mathrm{error}}$ measures variation within groups.
- Define the ratio $F = \mathrm{MS}_{\mathrm{groups}}/\mathrm{MS}_{\mathrm{error}}$ to be the F-statistic (named in honor of R. A. Fisher who developed ANOVA among many other accomplishments).
- When $H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_k$ is true (and the assumption of equal variances is also true), then both $\mathrm{E}(\mathrm{MS}_{\mathrm{groups}}) = \sigma^2$ and $\mathrm{E}(\mathrm{MS}_{\mathrm{error}}) = \sigma^2$ and the value of F should then be close to 1.
- However, if the population mean are not all equal, then $\mathrm{E}(\mathrm{MS}_{\mathrm{groups}}) > \sigma^2$ and we expect F to be greater than one, perhaps by quite a bit.

The F Distribution

Definition

If W_1 and W_2 are independent χ^2 random variables with d_1 and d_2 degrees of freedom, then

$$F = \frac{W_1/d_1}{W_2/d_2}$$

has an F distribution with d_1 and d_2 degrees of freedom.

- The mean of the $F(d_1, d_2)$ distribution is $d_2/(d_2 - 2)$ provided that $d_2 > 2$.
- The F distributions have different shapes, depending on the degrees of freedom, but are typically unimodal and skewed right.
- The R function `pf()` finds areas to the left under F distribution and the R function `qf()` finds quantiles. These functions work just like `pt()` and `qt()` except that two degrees of freedom need to be specified.

Sampling Distribution

- If we have k independent random samples and:
  - the null hypothesis $H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_k$ is true;
  - all population variances are equal $\sigma_i^2 = \sigma^2$;
  - individual observations are normal, $Y_i \sim \mathrm{N}(\mu, \sigma)$;

  then,
  - $(k-1)\mathrm{MS}_{\mathrm{groups}}/\sigma^2 \sim \chi^2(k-1)$;
  - $(n-k)\mathrm{MS}_{\mathrm{error}}/\sigma^2 \sim \chi^2(n-k)$;
  - $\mathrm{MS}_{\mathrm{groups}}$ and $\mathrm{MS}_{\mathrm{error}}$ are independent;
- It follows that

$$F = \frac{\mathrm{MS}_{\mathrm{groups}}}{\mathrm{MS}_{\mathrm{error}}} \sim F(k-1, n-k)$$

ANOVA Table

- The F statistic is the test statistic for the hypothesis test $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ versus H_A: not all means are equal.
- The steps for computing F are often written in an ANOVA table with this form.

Source	df	Sum of Squares	Mean Square	F	P value
Groups	$k-1$	SS_{groups}	MS_{groups}	F	P
Error	$n-k$	SS_{error}	MS_{error}		
Total	$n-1$	SS_{total}			

Total Sum of Squares

- The total sum of squares is the sum of squared deviations around the grand mean.

$$SS_{total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

- It can be shown algebraically that

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{j=1}^{k} n_j(\bar{Y}_j - \bar{Y})^2 + \sum_{j=1}^{k}(n_j - 1)s_j^2$$

or

$$SS_{total} = SS_{groups} + SS_{error}$$

Return to the Cuckoo Example

- The function `lm()` fits linear models in R.
- The function `anova()` displays the ANOVA table for the fitted model.

```
> cuckoo.lm = lm(eggLength ~ hostSpecies, data = cuckoo)
> anova(cuckoo.lm)

Analysis of Variance Table

Response: eggLength
            Df Sum Sq Mean Sq F value    Pr(>F)
hostSpecies  5 42.940  8.5879  10.388 3.152e-08 ***
Residuals  114 94.248  0.8267
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

There is very strong evidence that the mean sizes of cuckoo bird eggs within populations that use different host species are different (one-way ANOVA, $F = 10.4$, $df = 5$ and 114, $P < 10^{-7}$). This is consistent with a biological explanation of adaptation in response to natural selection; host birds may be more likely to identify an egg as not their own and remove it from the nest if its size differs from the size of its own eggs.

Summary Statistics

- The table can also be constructed from summary statistics.
- Note for example that the mean square error in the ANOVA table is a weighted average of the sample variances.

Host Species	n	mean	sd	variance
HedgeSparrow	14	23.12	1.07	1.14
MeadowPipet	45	22.30	0.92	0.85
PiedWagtail	15	22.90	1.07	1.14
Robin	16	22.57	0.68	0.47
TreePipet	15	23.09	0.90	0.81
Wren	15	21.13	0.74	0.55

Example Calculations

- *Degrees of freedom depends only on sample sizes.*
- $14 + 45 + 15 + 16 + 15 + 15 = 120$ so there are 119 total degrees of freedom.
- There are $k = 6$ groups, so there are 5 degrees of freedom for group.
- The difference is 114 degrees of freedom for error (or residuals).
- MS_{error} is the weighted average of sample variances

$$MS_{error} = \frac{(13)(1.07)^2 + (44)(0.92)^2 + (14)(1.07)^2 + (15)(0.68)^2 + (14)(0.90)^2 + (14)(0.74)^2}{114}$$

$$\doteq 0.827$$

More Calculations

- The grand mean:

$$\frac{(14)(23.12) + (45)(22.30) + (15)(22.90) + (16)(22.57) + (15)(23.09) + (15)(21.13)}{120}$$

$$\doteq 22.46$$

- Group sum of squares:

$$(14)(23.12-22.46)^2+(45)(22.30-22.46)^2+(15)(22.90-22.46)^2+(16)(22.57-22.46)^2+(15)(23.09-22.46)^2+(15)(21.13-22.46)^2$$

$$\doteq 42.94$$

- You should know how to complete a partially filled ANOVA table and how to find entries from summary statistics.

Variance Explained

Definition

The *proportion of variability explained by the groups*, or R^2 *value*, is defined as

$$R^2 = \frac{SS_{groups}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}}$$

and takes on values between 0 and 1.

- In the cuckoo example, the proportion of the variance explained is $42.94/137.19 \doteq 0.31$.

Estimation

- The ANOVA analysis provides strong evidence that the populations of cuckoo birds that lay eggs in different species of host nests have, on average, eggs of different size.
- It is more challenging to say in what ways the mean egg lengths are different.
- Estimating the standard error for each difference is straightforward.
- Finding appropriate multipliers for those differences may depend on whether or not the researcher is examining a small number of predetermined differences, or if the researcher is exploring all possible pairwise differences.
- In the former case, a t-distribution multiplier is appropriate, except that the standard error is estimated from all samples, not just two.
- In the latter case, there are many approaches, none perfect.

Standard Error

- When estimating the difference between two population means, recall the standard error formula (assuming a common standard deviation σ for all populations)

$$\text{SE}(\bar{Y}_i - \bar{Y}_j) = \sigma \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

- In the two-sample method, we pooled the two sample variances to estimate σ with s_{pooled}.
- In ANOVA, the square root of the mean square error, $\sqrt{\text{MS}_{\text{error}}}$ pools the data from all samples to estimate the common σ.
- This is only sensible if the assumption of equal variances is sensible.

Example

- For the cuckoo data, we have this estimate for σ.

$$\sqrt{\text{MS}_{\text{error}}} \doteq \sqrt{0.827} \doteq 0.91$$

- With six groups, there are 15 different two-way comparisons between sample means.
- The standard errors are different and depend on the specific sample sizes.
- It can be useful to order the groups according to the size of the sample means.

Example

Population	Mean	n	Population	Mean	n	Difference	SE
MeadowPipet	22.30	45	Wren	21.13	15	1.17	0.27
Robin	22.57	16	Wren	21.13	15	1.45	0.33
PiedWagtail	22.90	15	Wren	21.13	15	1.77	0.33
TreePipet	23.09	15	Wren	21.13	15	1.96	0.33
HedgeSparrow	23.12	14	Wren	21.13	15	1.99	0.34
Robin	22.57	16	MeadowPipet	22.30	45	0.28	0.27
PiedWagtail	22.90	15	MeadowPipet	22.30	45	0.60	0.27
TreePipet	23.09	15	MeadowPipet	22.30	45	0.79	0.27
HedgeSparrow	23.12	14	MeadowPipet	22.30	45	0.82	0.28
PiedWagtail	22.90	15	Robin	22.57	16	0.33	0.33
TreePipet	23.09	15	Robin	22.57	16	0.52	0.33
HedgeSparrow	23.12	15	Robin	22.57	16	0.55	0.33
TreePipet	23.09	15	PiedWagtail	22.90	15	0.19	0.33
HedgeSparrow	23.12	14	PiedWagtail	22.90	15	0.22	0.34
HedgeSparrow	23.12	14	TreePipet	23.09	15	0.03	0.34

Confidence Intervals

- Each of the fifteen differences can be estimated with confidence by using a t-multiplier times the SE for the margin of error.
- The t-multiplier is based on the confidence level and the error degrees of freedom.
- In the example, for a 95% confidence interval, the multiplier would be $t^* = 1.98$.
- Each of the fifteen confidence intervals would be valid, but it would be incorrect to interpret with 95% confidence that *each of the fifteen confidence intervals contains the corresponding difference in means*.

95% Confidence Intervals

Population	Population	a	b
MeadowPipet	Wren	0.63	1.71
Robin	Wren	0.80	2.09
PiedWagtail	Wren	1.12	2.43
TreePipet	Wren	1.30	2.62
HedgeSparrow	Wren	1.32	2.66
Robin	MeadowPipet	-0.25	0.80
PiedWagtail	MeadowPipet	0.07	1.14
TreePipet	MeadowPipet	0.25	1.33
HedgeSparrow	MeadowPipet	0.27	1.37
PiedWagtail	Robin	-0.32	0.98
TreePipet	Robin	-0.13	1.16
HedgeSparrow	Robin	-0.11	1.21
TreePipet	PiedWagtail	-0.47	0.84
HedgeSparrow	PiedWagtail	-0.45	0.89
HedgeSparrow	TreePipet	-0.64	0.70

Shorter Summary

- Rather than reporting all pairwise confidence intervals for differences in population means, researchers often list the sample means with different letters for collections of groups that are not significantly different from one another.
- Sometmes these collections of groups overlap.
- Here is an example with the cuckoo data.

Wren[A]	MeadowPipet[B]	Robin[B,C]	PiedWagtail[C]	TreePipet[C]	HedgeSparrow[C]
21.13	22.30	22.57	22.90	23.09	23.12

- Here Wren is significantly smaller than everything.
- Meadow Pipet and Robin are significantly larger than Wren, and Meadow Pipet (but not Robin) is significantly smaller that the other three.
- The top four are not significantly different from each other.

Simultaneous confidence intervals

- If we want to be 95% confident that *all population mean differences* are contained in their intervals, we need to increase the size of the multipler.
- This issue is known as *multiple comparisons* in the statistics literature.
- The method described in the text, *Tukey's honestly significant difference (HSD)* is based on the sampling distribution of the difference between the largest and smallest sample means when the null distribution is true, but assumes equal sample sizes.
- Other methods use slightly smaller multipliers for other differences; for example, the multiplier for the difference between the first and second largest sample means would be smaller than that for the largest and smallest sample means.
- It suffices to know that if you care about adjusting for multiple comparisons, that the multipliers need to be larger than the t-multipliers and that there are many possible ways to accomplish this.

Tukey's HSD in R

- R contains the function `TukeyHSD()` which can be used on the output from `aov()` to apply Tukey's HSD method for simultaneous confidence intervals.
- The method adjusts for imbalance in sample size, but may not be accurate with large imbalances.

Cuckoo Data

```
-------------- file cuckoo.txt --------------
eggLength hostSpecies
19.65 MeadowPipet
20.05 MeadowPipet
20.65 MeadowPipet
20.85 MeadowPipet
21.65 MeadowPipet
...
21.45 Wren
22.05 Wren
22.05 Wren
22.05 Wren
22.25 Wren
-------------- end of file ----------------
```

Reading in the Data

- Here is code to read in the data.
- We also use the lattice function `reorder()` to order the populations from smallest to largest egg length instead of alphabetically.
- This reordering is not essential, but is useful.
- The command `with()` allows R to recognize the names hostSpecies and eggLength without the dollar sign.
- The `require()` function loads in lattice if not already loaded.

```
> cuckoo = read.table("cuckoo.txt", header = T)
> require(lattice)
> cuckoo$hostSpecies = with(cuckoo, reorder(hostSpecies,
+     eggLength))
```

Fitting the ANOVA model

- We greatly prefer using `lm()` instead of `aov()`, but `TukeyHSD()` requires the latter.

```
> fit = aov(eggLength ~ hostSpecies, data = cuckoo)
```

Tukey's HSD

```
> TukeyHSD(fit)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = eggLength ~ hostSpecies, data = cuckoo)

$hostSpecies
                             diff          lwr       upr     p adj
MeadowPipet-Wren        1.16888889  0.383069115 1.954709 0.0004861
Robin-Wren              1.44500000  0.497728567 2.392271 0.0003183
PiedWagtail-Wren        1.77333333  0.810904595 2.735762 0.0000070
TreePipet-Wren          1.96000000  0.997571262 2.922429 0.0000006
HedgeSparrow-Wren       1.99142857  1.011964373 2.970893 0.0000006
Robin-MeadowPipet       0.27611111 -0.491069969 1.043292 0.9021876
PiedWagtail-MeadowPipet 0.60444444 -0.181375330 1.390264 0.2324603
TreePipet-MeadowPipet   0.79111111  0.005291337 1.576931 0.0474619
HedgeSparrow-MeadowPipet 0.82253968 0.015945760 1.629134 0.0428621
PiedWagtail-Robin       0.32833333 -0.618938100 1.275605 0.9155004
TreePipet-Robin         0.51500000 -0.432271433 1.462271 0.6159630
HedgeSparrow-Robin      0.54642857 -0.418146053 1.511003 0.5726153
TreePipet-PiedWagtail   0.18666667 -0.775762072 1.149095 0.9932186
HedgeSparrow-PiedWagtail 0.21809524 -0.761368960 1.197559 0.9872190
HedgeSparrow-TreePipet  0.03142857 -0.948035627 1.010893 0.9999990
```

Comparison

Population	Population	t-method		Tukey HSD	
		a	b	a	b
MeadowPipet	Wren	0.63	1.71	0.38	1.95
Robin	Wren	0.80	2.09	0.50	2.39
PiedWagtail	Wren	1.12	2.43	0.81	2.74
TreePipet	Wren	1.30	2.62	1.00	2.92
HedgeSparrow	Wren	1.32	2.66	1.01	2.97
Robin	MeadowPipet	-0.25	0.80	-0.49	1.04
PiedWagtail	MeadowPipet	0.07	1.14	-0.18	1.39
TreePipet	MeadowPipet	0.25	1.33	0.01	1.58
HedgeSparrow	MeadowPipet	0.27	1.37	0.02	1.63
PiedWagtail	Robin	-0.32	0.98	-0.62	1.28
TreePipet	Robin	-0.13	1.16	-0.43	1.46
HedgeSparrow	Robin	-0.11	1.21	-0.42	1.51
TreePipet	PiedWagtail	-0.47	0.84	-0.78	1.15
HedgeSparrow	PiedWagtail	-0.45	0.89	-0.76	1.20
HedgeSparrow	TreePipet	-0.64	0.70	-0.95	1.01

Interpretation

- There is evidence that the population mean length of cuckoo bird eggs in wren nests is smaller than those of all other cuckoo bird populations.
- Other comparisons are difficult to interpret, as we are not confident in the order of means, even though we are confident about some differences.
- Note that the Tukey confidence intervals are noticeably wider.

What you should know so far

You should know:

- how to complete a partially completed ANOVA table;
- how to fill an ANOVA table from summary statistics;
- how to find the pooled estimate of the common standard deviation;
- how to construct a confidence interval for the difference in two population means;
- why there may be a need to use a different method when contructing simutaneous confidence intervals.

Strontium in Trout Eggs

Case Study

- Researchers sampled brown trout eggs from six tributaries of the Taieri River in New Zealand.
- The strontium level from each egg sample was measured.
- The belief is that strontium levels will be high (average above 125 nmoles per gram dry weight eggs) in fish populations that spend significant time in the ocean, and be lower when the fish are a mixture between those that spend time in the ocean and are resident entirely in the rivers.
- Analyze the strontium levels in this data.

Strontium Levels
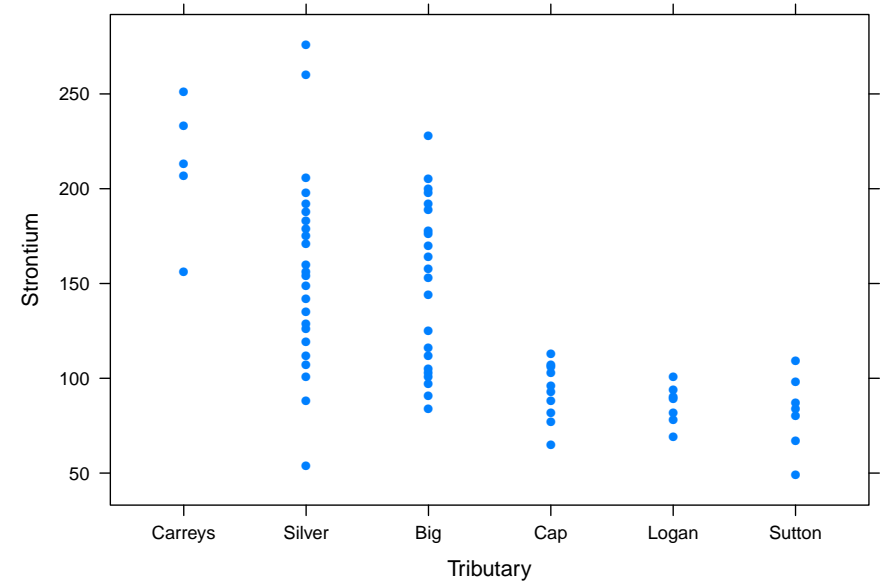
Data Summaries

```
> with(trout, sapply(split(Strontium, Tributary), mean))

  Carreys     Silver        Big        Cap      Logan     Sutton
212.00000 156.66667 149.50000   93.00000   86.14286   82.00000

> with(trout, sapply(split(Strontium, Tributary), sd))

 Carreys    Silver       Big       Cap     Logan    Sutton
35.79106 50.95921 43.56905 15.12907 10.66815 19.71463

> with(trout, sapply(split(Strontium, Tributary), length))

Carreys  Silver     Big     Cap   Logan  Sutton
      5      24      22      10       7       7
```
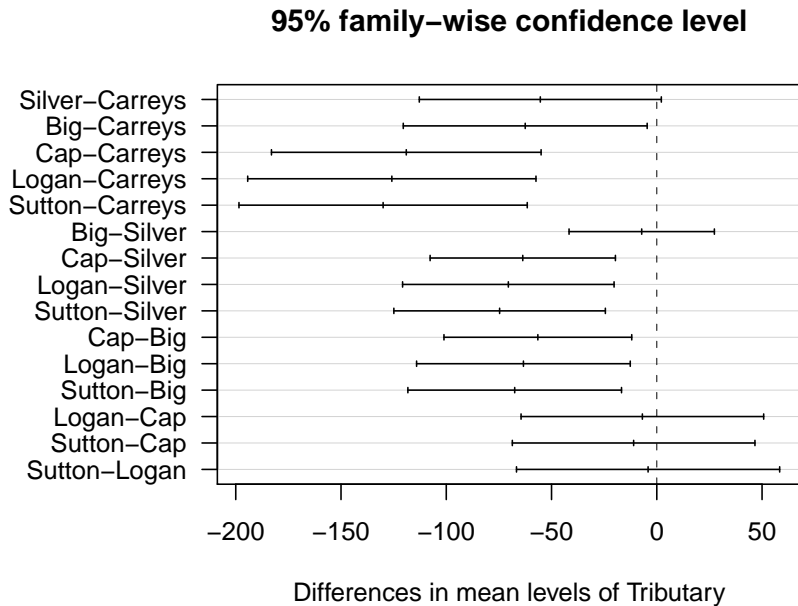
ANOVA for Trout Data

```
> trout.lm = lm(Strontium ~ Tributary, trout)
> anova(trout.lm)

Analysis of Variance Table

Response: Strontium
          Df Sum Sq Mean Sq F value    Pr(>F)
Tributary  5  99439 19887.8  12.499 1.201e-08 ***
Residuals 69 109790  1591.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Plot of Tukey HSD Intervals

95% family–wise confidence level

Differences in mean levels of Tributary

Linear Models

- ANOVA is an example of a *linear model*.
- In a linear model, a response variable Y is modeled as a mean plus error, where
  - the mean is a *linear function* of parameters and covariates;
  - the error is random normally distributed mean-zero variation.
- A linear function takes the form

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where the $\{\beta_i\}$ are parameters and the $\{x_i\}$ are covariates.

Linear Model

- A linear model takes the following form.

$$Y_i = \mu_{j(i)} + \varepsilon_i$$

where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ and $\mu_{j(i)} = \mathrm{E}(Y_i)$.
- There are multiple ways to parameterize a one-way ANOVA model.
- Consider a toy example with $k = 3$ groups with means 16, 20, and 21.

First Parameterization

- One way to parameterize a one-way ANOVA model is to treat one group as a reference, and parameterize differences between the means of other groups and the reference group.
- If the first group is selected as the reference:
  - $\beta_0 = \mu_1$;
  - $\beta_1 = \mu_2 - \mu_1$;
  - $\beta_2 = \mu_3 - \mu_1$.
- Using the example $\mu_1 = 16$, $\mu_2 = 20$, and $\mu_3 = 21$, we have $\beta_0 = 16$, $\beta_1 = 4$ and $\beta_2 = 5$.
- Notice that the statement
  the first mean is 16, the second mean is four larger than the first, and the third mean is five larger than the first

  is just a different way to convey the same information as
  the first mean is 16, the second is 20, and the third is 21.

First Parameterization (cont.)

- Define these covariates (here, indicator random variables):
  - x_{1i} is 1 if the ith observation is in group 2 and be 0 if it is not.
  - x_{2i} be 1 if the ith observation is in group 3 and be 0 if it is not.
- Then,
$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$
- In this example the three means $\{\mu_j\}$ are reparameterized with three parameters $\{\beta_j\}$.
- Notice:
  - if the ith observation is in group 1, then $x_{1i} = 0$ and $x_{2i} = 0$ so $Y_i = \beta_0 + \varepsilon_i$;
  - if the ith observation is in group 2, then $x_{1i} = 1$ and $x_{2i} = 0$ so $Y_i = \beta_0 + \beta_1 + \varepsilon_i$;
  - if the ith observation is in group 3, then $x_{1i} = 0$ and $x_{2i} = 1$ so $Y_i = \beta_0 + \beta_2 + \varepsilon_i$.

lm() in R

- The previous parameterization is the default in R.
- Consider the cuckoo example again.

```
> cuckoo.lm = lm(eggLength ~ hostSpecies, data = cuckoo)
> summary(cuckoo.lm)

Call:
lm(formula = eggLength ~ hostSpecies, data = cuckoo)

Residuals:
    Min      1Q  Median      3Q     Max
-2.64889 -0.44889 -0.04889  0.55111  2.15111

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             21.1300     0.2348  90.004  < 2e-16 ***
hostSpeciesMeadowPipet   1.1689     0.2711   4.312 3.46e-05 ***
hostSpeciesRobin         1.4450     0.3268   4.422 2.25e-05 ***
hostSpeciesPiedWagtail   1.7733     0.3320   5.341 4.78e-07 ***
hostSpeciesTreePipet     1.9600     0.3320   5.903 3.74e-08 ***
hostSpeciesHedgeSparrow  1.9914     0.3379   5.894 3.91e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9093 on 114 degrees of freedom
Multiple R-squared: 0.313,  Adjusted R-squared: 0.2829
F-statistic: 10.39 on 5 and 114 DF,  p-value: 3.152e-08
```

lm() in R (cont.)

- In this formulation, the mean length of cuckoo birds laid in wren nests is the intercept β_0.
- This is estimated as 21.13, the wren group mean.
- The other parameters are differences between means of other groups and means of the wren group.
- For example, the meadow pipet mean group mean is 22.30, or 1.17 larger than the wren group mean.
- The summary contains inferences for six parameters.
- The first line tests $H_0 \colon \beta_0 = 0$, or that the mean length of the eggs in the wren group is zero. This is biologically meaningless and overwhelmingly rejected.
- Each other row one of the pairwise comparisions between the wren group and the others.
- Each p-value is (much) less than 0.05, consistent with the 95% confidence intervals for these differences not containing 0.
- None of the other ten pairwise comparisons is shown, though.

Confidence Intervals from the Summary

- We can construct some confidence intervals for population mean differences from this summary.
- The residual error 0.9093 on 114 degrees of freedom matches $\sqrt{0.8267}$ from the ANOVA table.
- The standard error for the meadow pipet minus wren group mean difference is 0.2711 which matches
$$0.9093 \times \sqrt{\frac{1}{15} + \frac{1}{45}}$$
- The critical t quantile with 114 degrees of freedom for a 95% confidence interval is 1.98, so the margin of error is 0.54.
- Adding and subtracting this to the difference 1.17 results in the 95% confidence interval
$$0.63 < \mu_{\text{meadow pipet}} - \mu_{\text{wren}} < 1.71$$
- This matches the result from an earlier slide.
- This interval (and the others) do not compensate for multiple comparisons.

Quick summary

- A one-way analysis of variance model is fit in R using `lm()`.
- The results of this model fit can be summarized using `anova()` which displays an ANOVA table.
- The ANOVA table is a structured calculation of a test statistic for the null hypothesis $H_0 \colon \mu_1 = \cdots = \mu_k$ with an F test.
- The results can also be summarized with `summary()` which displays estimated coefficients and standard errors for model parameters and t-tests for the hypotheses $H_0 \colon \beta_j = 0$.
- The model parameters include $k - 1$ of the pairwise differences, but not all of them.
- Standard errors for other differences may be found by hand $\hat{\sigma}\sqrt{1/n_i + 1/n_j}$ or by changing the order of the levels in the factor.

Cautions and Concerns

- One-way ANOVA assumes independent random sampling from different populations.
- The F-distribution of the test statistic assumes equal variances among populations and normality:
  - ▶ if not, the true sampling distribution is not exactly F;
  - ▶ However, the method is robust to moderate deviations from equal variance;
  - ▶ and, the method is robust to moderate deviations from normality.
  - ▶ If the equal variance or normal assumptions (or both) are untenable, then the p-value could be found from the null distribution of the F statistic from a randomization test where groups are assigned in their given sizes at random.

Extensions

- Linear models can be extended by adding additional explanatory variables.
- If all explanatory variables are factors, then the model is mutli-way ANOVA.
- If all explanatory variables are quantitative, then the model is regression.
- If the levels of a factor are considered as *random draws from a population* instead of unknown fixed parameters, then the model is called a *random effects model*.
- Models with two or more explanatory variables can include parameters for *interactions*.
- If the response variable is not normal (or transformable to normal) and another distribution is more appropriate (such as binomial or Poisson), then we should consider instead a *generalized linear model*.