TREATMENT COMPARISONS

Relating Fisher's 1926 paper to the separate-means model

Recall the separate-means model (cell-mean parameterization):

$$y_{ij} = \mu_i + e_{ij} \quad (i = 1, \ldots, t; j = 1, \ldots, r),$$

where e_{ij}'s are assumed i.i.d. normal $N(0, \sigma^2)$.

- To compare population means from two groups, say μ_1 and μ_2, we will rely on the sample means $\overline{y}_{1\cdot}$ and $\overline{y}_{2\cdot}$.
- To test the statistical significance of the difference, we compare $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ to its standard error. We need to estimate the error in the difference between units treated differently. However, our estimate of the (standard) error will be based on differences between units treated alike (see the expression of the SSE and MSE).
- For this error estimation to be valid for comparing two group means, any two units chosen from the same group must be no more and no less similar than any two units chosen from different groups.
- We can satisfy this requirement by assigning experimental units to different treatment groups completely at random.

Recall the separate-means model with the cell-mean parameterization, where each observation is represented as the group mean plus an error term. The error terms are assumed iid normal with mean 0 and variance sigma squared.

To compare population means of the two groups, say mu 1 and mu 2, we will rely on the two corresponding group sample means y bar 1 dot and y bar 2 dot.

To test the statistical significance of the difference, we compare the difference of the two group sample means, y bar 1 dot minus y bar 2 dot, to its standard error.

We need to estimate the error in the difference between units treated differently. However, our estimate of the (standard) error will be based on differences between units treated alike: we can see this from the expression of the SSE and MSE.

For this error estimation to be valid for comparing two group means, any two units chosen from the same group must be no more and no less similar than any two units chosen from different groups.

We can satisfy this requirement by assigning experimental units to different treatment groups completely at random. One can also say that we assign treatments to experimental units completely at random.

A note on the i.i.d assumption

- i.i.d. means “independent and identically distributed”. We often have an i.i.d. and normal assumption on the error terms in many of the models we explore. The i.i.d normal assumption on the error terms (e_{ij}) simplifies the mathematical analysis of the SSE, MSE, the F-test in the ANOVA, t-test for contrasts (today) and so on.
- The CRD assumption—experimental units are completely randomly assigned to treatments—is not exactly equivalent to the i.i.d. normal assumption, but most results we derive under the i.i.d. normal assumptions will still hold (approximately) under the CRD assumption.
- So when we see the i.i.d. assumption, we should understand that the results that follow can be justified by randomization.
- When analyzing a designed experiment, we thus need to pay attention to the details of the randomization scheme. For example, If the randomization is restricted to a subgroup, then we can only assume that the error terms are i.i.d. for that subgroup.
- The key is to “follow the randomization”. In a completely randomized experiment, all error terms can be assumed i.i.d..

A note on the iid assumption

i.i.d. means “independent and identically distributed”. We often have an i.i.d. and normal assumption on the error terms in many of the models we explore. The i.i.d. normal assumption on the error terms simplifies the mathematical analysis of the SSE, MSE, the F-test in the ANOVA, t-test for contrasts (which we will talk about today) and so on.

The CRD assumption—experimental units are completely randomly assigned to treatments—is not exactly equivalent to the i.i.d. normal assumption, but most results we derive under the i.i.d. normal assumptions will still hold (approximately) under the CRD assumption.

So when we see the i.i.d. assumption, we should understand that the results that follow can be justified by randomization.

When analyzing a designed experiment, we thus need to pay attention to the details of the randomization scheme. For example, If the randomization is restricted to a subgroup, then we can only assume that the error terms are i.i.d. for that subgroup.

The key is to “follow the randomization”. In a completely randomized experiment, all error terms can be assumed i.i.d..

The i.i.d. assumption is more essential than the normal assumption. Due to the central limit theorem, the sample means will almost always be approximately normally distributed. Therefore, most of the results about the mean parameters will hold approximately for non-normal data if the sample size is not too small.

Comparing two group means

Suppose we have a statistically significant result from a single-factor ANOVA. It suggests not all treatment means (group population means) are the same, but it doesn't tell us which treatment means differ from which others.

How do we compare two group means? Say μ_1 and μ_2.

In the blood pressure example, the corresponding research question is comparing the mean change in blood pressure in the two dose groups (say dose 20 versus dose 0).

We can use a two-sample t-test: we take the difference of the two group means and see how many times greater it is than the **estimated standard error**.

Suppose we have a statistically significant result from a single-factor ANOVA. It suggests not all treatment means (group population means) are the same, but it doesn't tell us which treatment means differ from which others.

How do we compare two group means? Say mu 1 and mu 2.

In the blood pressure example, the corresponding research question is comparing the mean change in blood pressure in the two dose groups (say dose 20 versus dose 0).

We can use a two-sample t-test: we take the difference of the two group sample means and see how many times greater it is than the estimated standard error.

Comparing the group means

In the blood pressure example,

- $\overline{y}_{2\cdot} - \overline{y}_{1\cdot} = -4.2 - 3 = -7.2$

and the estimated standard error of this difference is

- $\sqrt{\frac{2}{r}MSE} = \sqrt{\frac{2}{5} \times 42.47} = 4.12$

The t-statistic is the ratio of two

- $t = \frac{\overline{y}_{2\cdot} - \overline{y}_{1\cdot}}{\sqrt{\frac{2}{r}MSE}} = \frac{-7.2}{4.12} = -1.75$

Comparing this t-statistic to its distribution under null, which is a t-distribution with $N - t = 12$ d.f., gives a p-value of 0.106.

Note: The d.f. for the t-test is same as the d.f. for the MSE used for estimating σ^2. We get the MSE and its d.f. from the ANOVA table.

(See *script1* for how to do the calculations in R.)

In the blood pressure example, the difference of the two group sample means is -7.2.

The estimated standard error of this difference is the square root of 2 over r times the MSE, and equals 4.12.

The t-statistic for comparing the two group population means is the ratio of the group sample mean difference over its estimated standard error. It equals -1.75.

Comparing this t-statistic to its null distribution, which is a t-distribution with N - t = 12 degrees of freedom, gives a p-value of 0.106. There is no evidence that the mean changes in blood pressure are significantly different between the two dose groups.

The number of d.f. for the t-test is same as the number of d.f. for the MSE used for estimating sigma squared. We can get the MSE and its d.f. from the ANOVA table.

More general contrasts among means

A general contrast is defined as

$$C = k_1\mu_1 + k_2\mu_2 + \cdots + k_t\mu_t = \sum_{i=1}^{t} k_i\,\mu_i,$$

where the k's are constants, and $\sum_{i=1}^{t} k_i = 0$.

For example,

- $C = \mu_2 - \mu_1$ is contrast with $k_1 = 1$ and $k_2 = -1$.
- $C = \frac{\mu_1+\mu_2}{2} - \mu_3$ is a contrast, with $k_1 = k_2 = 1/2$ and $k_3 = -1$.

We often use a contrast when stating a null hypothesis: for example, we can construct C such that $C = 0$ correspond to the null hypothesis of interest.

A general contrast is linear combination of the group mean parameters when the sum of the coefficients are 0.

For example, C equals mu 2 minus mu 1 is contrast with k 1 equals 1 and k 2 equals -1. C equals mu 1 plus mu 2 divided by 2 minus mu 3 is a contrast. In this expression, k 1 equals k 2 equals one half and k 3 equals -1.

We often use a contrast when stating a null hypothesis: for example, we can construct the contrast C such that C equal 0 corresponds to the null hypothesis of interest.

Estimate a contrast, confidence interval, and hypothesis test

1. We estimate a contrast $C = \sum_{i=1}^{t} k_i \mu_i$ using $c = \sum_{i=1}^{t} k_i \overline{y}_{i\cdot}$, where $\overline{y}_{i\cdot} = \sum_{j=1}^{r_i} y_{ij} / r_i$ is the ith group sample mean. The estimated variance of c is

$$s_c^2 = s^2 \sum_{i=1}^{t} \frac{k_i^2}{r_i}.$$

where s^2 is the MSE from fitting the separate-means model.

2. A $100(1-\alpha)\%$ confidence interval for a contrast C is given by

$$c \pm t_{\alpha/2, N-t} \cdot s_c$$

where $t_{\alpha/2,N-t}$ (called **the critical value**) is the $(1-\alpha/2)$th quantile of a t distribution with $N-t$ d.f..

3. We can test the null hypothesis that $C = 0$ by comparing c/s_c to a t-distribution with $N-t$ d.f. (usually a two-sided test).

To estimate a contrast, we simply replace the group population means in the expression by group sample means. We can use little c to denote this statistic.

Note, again, that big C is a population parameter and little c is a statistic.

The variance of little c can be estimated by s squared times the sum of k i squared divided by r i over all groups, where s squared is the MSE from fitting a separate-means model to the data set.

To remember this formula: note that s squared over r i estimate the variance of y i bar and recall that when taking variance of a linear combination of independent random variables, we can take the variance term by term, but we need square the coefficient.

The estimated standard error of little c is then the square root of the estimated variance.

A (1-alpha) confidence interval for the contrast C is given by little c plus or minus a critical value times the estimated standard error.

The critical value, (t alpha over 2, N minus t), is the (1-alpha/2)th quantile a t distribution with N -t degrees of freedom. That is, the area under the density curve to the right of this number is alpha over 2.

We can test the null hypothesis big C equals o by comparing the little c divided by its estimated standard error to a t distribution with N minus t degrees of freedom.

Usually, we will do a two-sided test.

Example. Cumulative GPA of 90 OSU students by residency (data set `resgpa`).

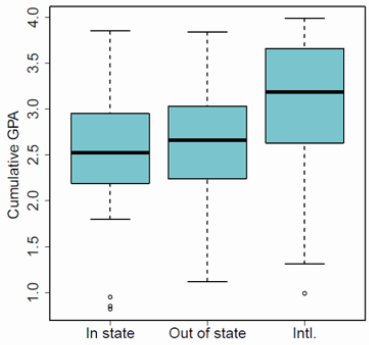
	Oregon (1)	Out of state (2)	International (3)
r_i	30	30	30
$\bar{y}_{i\bullet}$	2.49	2.63	3.07
s_i	0.770	0.592	0.723

Now let's look at an example.

The resgpa data set consists of residency and GPA information for 90 students randomly selected from a large population of OSU students that graduated in the 1990's. The sampling was restricted to yield equal numbers of students (30) in each residency group.

This table summaries the group size, the group sample mean and standard deviation for each group.

Boxplot and the ANOVA table for the GPA data



Single-factor ANOVA reveals strong evidence that the mean cumulative GPA varies among residency groups:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residency	2	5.581	2.791	5.7077	0.004686
Residuals	87	42.537	0.489		

Single-factor ANOVA reveals strong evidence that the mean cumulative GPA varies among residency groups. From the boxplot we can see, for example, that the average GPA is much higher for international students.

Compare two group means

We could do pairwise comparisons among the means, e.g., comparing group 1 and 3 using the contrast $C = \mu_1 - \mu_3$ ($k_1 = 1$, $k_2 = 0$, and $k_3 = -1$).

Exercise. Compute c and s_c according to the formulas we discussed on page 7.

We could do pairwise comparisons among the means.

As an exercise, suppose that we want to compare group 1 and group 3 using the contrast C equals mu 1 minus mu 3. Compute the little c and its estimated standard error according to the formulas we discussed earlier using information in the mean table and the ANOVA table.

Contrast involving three means

Compare the population mean GPA among international students to that among domestic students: we can compare the population mean GPA in group 3 (international students) to the average of the means in groups 1 and 2 (domestic students), and examine the contrast $C = \frac{\mu_1+\mu_2}{2} - \mu_3$,

- Estimate C by $c = \frac{\overline{y}_{1\cdot}+\overline{y}_{2\cdot}}{2} - \overline{y}_{3\cdot} = \frac{2.49+2.63}{2} - 3.07 = 0.51.$
- $s_c = \sqrt{MSE \cdot \sum_{i=1}^{3} \frac{k_i^2}{r_i}} = \sqrt{0.489[(1/2)^2/30 + (1/2)^2/30 + (-1)^2/30]} = 0.156.$

For testing $H_0: \frac{\mu_1+\mu_2}{2} - \mu_3 = 0$, compare c/s_c to a t-distribution with 87 d.f.:

```
pt(0.51/0.156, df=87, lower.tail=FALSE) *2
## [1] 0.001545623
```

The p-value $p = 0.0015$. We have strong evidence of a difference.

(*Note:* The MSE and its number of d.f. are from the ANOVA table. The number of d.f. of the t-test is the same as the number of d.f. for the MSE.)

Now suppose we want to compare the population mean GPA among international students to that among domestic students.

We can compare the population mean GPA in group 3 to the average of the means in groups 1 and 2, and examine the contrast (mu 1 plus mu 2) divided by 2 minus mu 3.

As discussed earlier, to estimate this contrast, we simply replace the population mean parameters by corresponding group sample means.

The formula for the standard error for little c gives s_c equal 0.156. Note the we can get the MSE from the ANOVA table.

For testing the null hypothesis that the population average of group 1 and group 2 mean population GPA is the same as the mean population GPA in group 3, we compare c divided by s_c to a t distribution with 87 d.f..

We can use the R code listed here to find out the p-value.

The p-value is 0.0015. We have strong evidence of a difference.

Note that the number of d.f. for the t-test is the same as the number of d.f. for the MSE, which we can get from the ANOVA table.

Confidence intervals

If $\overline{y}$ is the mean of n observations, then $Var(\overline{y}) = Var(y)/n$.

Since $\overline{y}_{i\cdot}$ is the mean of r observations from a distribution having variance σ^2, we have $Var(\overline{y}_{i\cdot}) = \sigma^2/r$, which we can estimated by MSE/r.

A $100(1-\alpha)\%$ confidence interval for μ_i is given by

$$\overline{y}_{i\cdot} \pm t_{\alpha/2,N-t} \cdot \sqrt{\frac{MSE}{r}}.$$

where the critical value $t_{\alpha/2,N-t}$ is the $(1-\alpha/2)$th quantile of the t-distribution with $N-t$ degrees of freedom.

In R, we can use the `qt` function to compute $t_{\alpha/2,N-t}$ as follows:

```
qt(1 - alpha/2, N-t);
qt(alpha/2, N-t, lower.tail=FALSE); ## This is more accurate for extremely small
alpha.
```

For an i.i.d. sample, the variance of the sample mean is the variance of a single observation divided by the sample size. Intuitively, we know that taking average of n independent random variables will reduce the variance.

The same formula applies to the group sample means with sample sizes being the group sizes. So the variance of y bar i dot is the variance of a single observation divided by r, which can be estimated by MSE from the fitted separate-means model divided by r.

A (1 minus alpha) confidence interval for mu i is given by the group sample mean plus or minus a critical value times the standard error of the group sample mean.

The standard error of (y bar i dot) is the square root of its estimated variance, which is the MSE divided by r.

The critical value, (t alpha over 2 N-t), is the (1-alpha over 2)-th quantile of the t-distribution with N-t degrees of freedom. The area under the t density curve to the left of this point is 1-alpha/2. Equivalently, The area under the t density curve to the right of this point, that is, the area of the upper tail, is alpha over 2.

In R, this t quantile can be computed using the qt function. Here, I show two function calls to qt. They are usually equivalent. The second one is more accurate when alpha is extremely small.

Example. Blood pressure data.

Find a 95% confidence interval for the mean change in blood pressure in the placebo group (group 1).

$$\overline{y}_{1\cdot} = 3.0, r = 5$$

$$MSE = 42.5, df = N - t = 15 - 3 = 12$$

$$SE_{\overline{y}_{1\cdot}} = \sqrt{\frac{MSE}{r}} = \sqrt{\frac{42.5}{5}}$$

$$t_{0.025,12} = 2.179$$

To find this t-distribution quantile in R, use

```
qt(0.975, 12);
## [1] 2.178813
```

The 95% confidence interval is thus $3.0 \pm 2.179 \cdot \sqrt{42.5/5} = (-3.4, 9.4)$.

Now let's look at an example. We will use the blood pressure data again.

Suppose we want to find a 95 percent confidence interval for the mean change in blood pressure in the placebo group (that is, group 1).

From the group sample mean table and the ANOVA table, we can find values of the group sample mean, the group size, the MSE from the fitted separate-means model, and its number of degrees of freedom.

With these values, we can compute the confidence interval: the confidence interval has the form of a group sample mean plus or minus a t-critical value times the estimated standard error of the sample mean.

The standard error depends on the MSE and the group size.

The critical value depends on the significance level and the number of d.f. of the standard error.

The resulting 95 percent confidence interval is (-3.4, 9.4).

To find this t-distribution quantile, we can use the qt function in R.

Discussion: an alternative confidence interval

In the last example, the standard error of $\overline{y}_{1\cdot}$ is estimated as $\sqrt{\frac{MSE}{r}}$ (which estimates $\sqrt{\frac{\sigma^2}{r}}$).

In stead of using the MSE, which is a pooled estimation of σ^2, one could also estimate σ^2 using the sample variance s_1^2 of observations from group 1.

And one could construct the confidence interval based on the variance of just those observations in the zero-dose group. When might this be a good idea?

Then the resulting 95% confidence interval is:

$$\overline{y}_{1\cdot} \pm t_{0.025,4} \cdot \sqrt{\frac{s_1^2}{5}} = 3.0 \pm 2.78 \cdot \sqrt{8.2^2/5} = (-7.2, 13.2).$$

Compare this confidence interval to the one in the last example. Comments on their differences. Which of these confidence intervals is better?

In the last example, the standard error of (y bar 1 dot) is estimated as (the square root of the MSE divided by r), which estimates (the square root of sigma squared by r).

In stead of using the MSE, which is a pooled estimation of (sigma squared), one could also estimate (sigma squared) using the sample variance, (s 1 squared), of observations from group 1, the zero-dose group.

And one could construct the confidence interval based on the variance of just those observations in the zero-dose group. When might this be a good idea?

Then the resulting 95 percent confidence interval is given here in this expression.

Compare this confidence interval to the one in the last example. Comments on their differences. Which of these confidence intervals is better?

Note on the alternative confidence interval

Comparing to the previous confidence interval, this one is wider.

Since we only used information from a single group to estimate the variance, there is more uncertainty in the estimate.

Note that the d.f. for s_1^2 is 4 while the d.f. for the MSE from the fitted separate-means model is 12.

Using MSE to estimate σ^2 assumes that all groups have equal variance.

If there is evidence that the groups do not have equal variance. The confidence interval based on s_1^2 is more reliable—in terms of having the correct coverage probability—even though it might be wider.

Comparing to the previous confidence interval, this one is wider.

Since we only used information from a single group to estimate the variance, there is more uncertainty in the estimate.

Note that the d.f. for (s 1 squared) is 4 while the d.f. for the MSE from the fitted separate-means model is 12.

Using MSE to estimate (sigma squared) assumes that all groups have equal variance.

If there is evidence that the groups do not have equal variance. The confidence interval based on (s 1 squared) is more reliable—in terms of having the correct coverage probability—even though it might be wider.