STAT 515 Lec 17

Comparative experiments and analysis of variance

Karl Gregory

Comparative experiments and analysis of variance

Much of the content in this section was developed for the first time by the British statistician and biologist Ronald A. Fisher in the early 1900s.

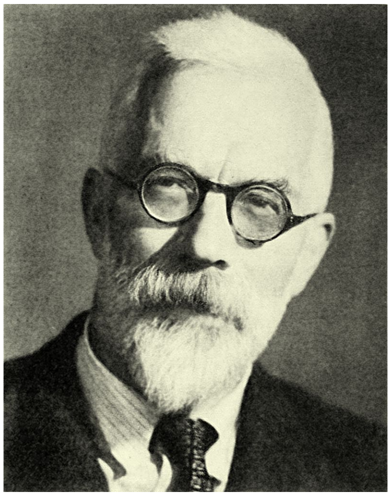


Figure 1: Ronald A. Fisher (1890) – (1962)

We will first draw a distinction between two types of studies which we shall call *comparative experiments* and *comparative observational studies*.

> *Comparative experiments* place subjects under different conditions through random assignment and compare outcomes.
>
> - Randomly assign plant clones to different drought conditions and compare $C0_2$ uptake.
> - Randomly assign tracts of a field to different fertilizer treatments and compare yields.

  - Randomly assign half of a group of rats to a period of radon exposure, keeping the other half unexposed, and compare metrics of carcinogenesis between the groups.

  Comparative observational studies compare subjects with different properties or which exist under different circumstances.

  - Compare development metrics of children from different socio-economic backgrounds.
  - Compare attitudes toward recycling of college students in Columbia and Greenville.

The cardinal difference between an experiment and an observational study is that in an experiment the investigator randomly assigns subjects to different conditions, whereas in an observational study the investigator does not assign subjects to conditions, but observes the subjects without changing their circumstances.

Observational studies are beset with the problem of *confounding variables.* A *confounding variable* is any unrecorded property or circumstance of the subjects which is associated with the outcome of interest as well as with the any of the properties or circumstances of the subjects which are recorded in the study. For example, an observational study may attempt to relate children's school grades (the outcome of interest) to the income of their parents, but there may be no information available about alcoholism in the children's homes; if the rate of alcoholism in the home is related to income as well as to children's grades in school, then alcoholism in the home is a confounding variable. If this is the case, a study which measures only children's grades and the income of their parents cannot attribute lower grades to lower parental income, even if the data show that children whose parents earn less make poorer grades.

Confounding variables lurk in every observational study. Yes, in all of them. It is therefore inappropriate to draw conclusions of causality from any observational study. You cannot say, "Circumstance A causes outcome Y," on the basis of an observational study, because there may be an unrecorded circumstance B commonly occurring with A which is the true agent causing outcome Y. One cannot know. One can only say, "Outcome Y is associated with circumstance A," by which we mean that the two often occur together, but A may or may not exert a causal influence upon Y.

Experiments eliminate the problem of confounding variables through the random assignment of the subjects to different conditions. The process of random assignment disassociates all unrecorded properties or circumstances of the subjects from the conditions of interest in the study. After randomization, no property or circumstance of the subjects has an association with the conditions imposed by the investigator. As a result, experiments, in contrast to observational studies, are capable of yielding causal conclusions.

The rest of this lecture will concern comparative experiments.

A vocabulary for comparative experiments

Words to know are

- *Treatment*: Any condition imposed by the investigator is called a treatment.
- *Experimental unit (EU)*: we generically refer to each subject in the study, whether a person, object, animal, or other entity, as an experimental unit.
- *Response*: The outcome which we measure on each experimental unit after administering the treatment is called the response.

Example. This example is taken from [1]. Twelve steaks were randomly assigned (three each) to four different packaging conditions (Commercial, Vacuum, Mixed Gas, CO_2). After 9 days at 4° C, the number of bacteria per cm^2 over the surface of the steak was recorded. Of interest is whether "some form of controlled gas atmosphere would provide a more effective packaging environment [than commercial or vaccuum] for meat storage". The raw data are

Steak	Packaging	log(# bact/cm^2)
1	Commercial	7.66
6	Commercial	6.98
7	Commercial	7.80
12	Vacuum	5.26
5	Vacuum	5.44
3	Vacuum	5.80
10	Mixed Gas	7.41
9	Mixed Gas	7.33
2	Mixed Gas	7.04
8	CO_2	3.51
4	CO_2	2.91
11	CO_2	3.66

The experimental units are the steaks, the treatments are the four packaging types, and the response is the natural logarithm of the number of bacteria per cm^2 over the surface of each steak.

It is natural to compute the means of each treatment group and to compare them:

Packaging	mean of log(# bact/cm^2)
Commercial	7.48
Vaccuum	5.50
Mixed Gas	7.26
CO_2	3.36

From this table, it looks like the packaging with CO_2 achieved the least bacterial growth. However, we know by now that need to do a more rigorous analysis than just looking at the sample means. If we were to repeat this experiment with 12 different steaks, we would get different response values and different treatment means. The question is, to what extent do the treatment means differ because of true differences in the packaging methods and to what extent do they differ because of experimental variability? We address these questions in the following with a view to hypothesis testing.

The cell means model ("One-way ANOVA")

In statistics we often try to make sense of where our data come from by writing down a mathematical expression for each value. We call this expression a *model*, and it is supposed to describe the mechanism working in the background to produce the data we observe. Writing down a model for our data helps us to formulate testable hypotheses which match our research questions. For the data in Example , a model called the *cell means model* or the *one-way ANOVA model* is often posited as the mechanism through which the response values come to be.

Before we can write down the cell means model, we need to define some quantities pertaining to a comparative experiment:

- Let K denote the number of treatments.
- Let $n_1, \dots, n_K$ denote the numbers of EUs assigned to treatments $1, \dots, K$, respectively.
- Let $N = n_1 + \cdots + n_K$ denote the total number of EUs.
- let Y_{ij} denote the response of the jth experimental unit of the ith treatment group for $j = 1, \dots, n_i$ and $i = 1, \dots, K$.

Now we can define the cell means model.

Definition: Cell means or one-way ANOVA model

Under the cell means or one-way ANOVA model we have

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma_\varepsilon^2) \tag{1}$$

for $j = 1, \dots, n_i$ and $i = 1, \dots, K$, where

- μ_i is called the population mean for the ith treatment
- ε_{ij} is called an error term and represents the deviation of the response Y_{ij} from the population mean of treatment i.

By the expression $\varepsilon_{ij} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma_\varepsilon^2)$ it is meant that all the error terms ε_{ij} are independent from one another and each one has a Normal distribution centered at zero with variance σ_ε^2. The independence of the ε_{ij} means that the value of one does not affect the value of any other one.

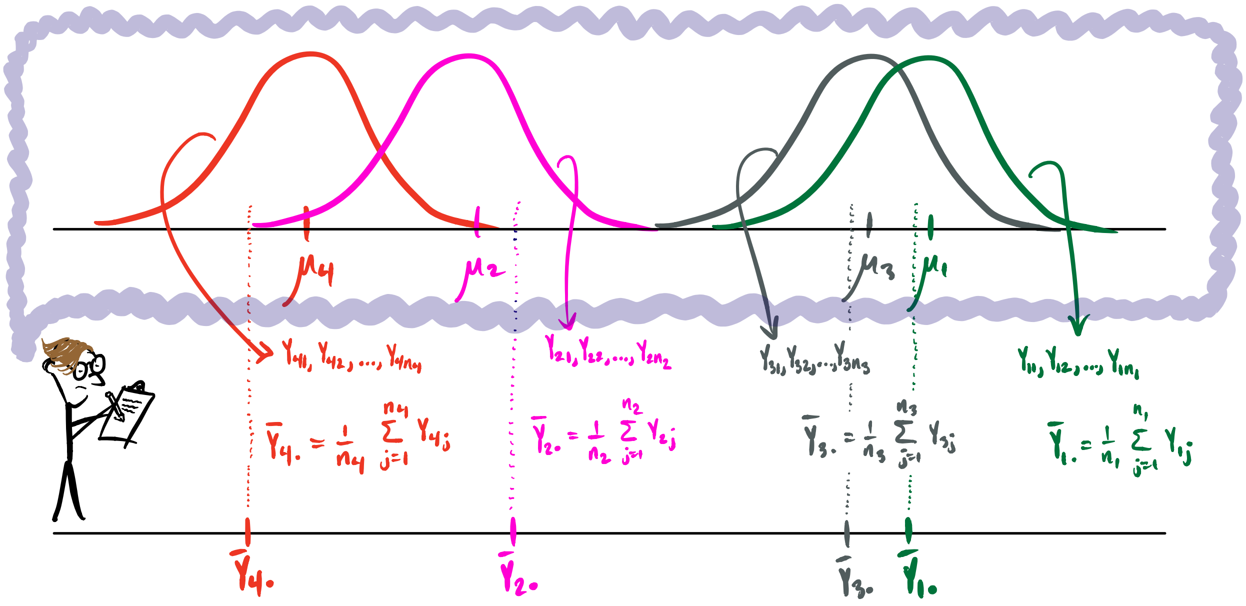
Under the cell means or one-way ANOVA model, it is typically of interest to estimate the population treatment means $\mu_1, \dots, \mu_K$ and to test hypotheses about them. Natural estimators of $\mu_1, \dots, \mu_K$ are given by

$$\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}, \quad i = 1, \dots, K,$$

which are the means of the responses in the treatment groups (placing a '.' in the subscript in the place of an index over which a sum has been taken is a common convention).

In Example with the steaks, we have $K = 4$ for the four packaging types, $n_1 = n_2 = n_3 = n_4 = 3$, since three steaks were assigned to each treatment, and the total number of experimental units is $N = n_1 + n_2 + n_3 + n_4 = 12$. We then have the responses $Y_{11} = 7.66, Y_{12} = 6.98, \dots, Y_{43} = 3.66$. The means of the treatment groups are $\bar{Y}_{1.} = 7.48$, $\bar{Y}_{2.} = 5.50$, $\bar{Y}_{3.} = 7.26$, and $\bar{Y}_{4.} = 3.36$.

We may depict a cell means model with $K = 4$ treatments as


μ_4
μ_2
μ_3
μ_1
$Y_{41}, Y_{42}, \dots, Y_{4n_4}$
$Y_{21}, Y_{22}, \dots, Y_{2n_2}$
$Y_{31}, Y_{32}, \dots, Y_{3n_3}$
$Y_{11}, Y_{12}, \dots, Y_{1n_1}$
$\bar{Y}_{4.} = \frac{1}{n_4} \sum_{j=1}^{n_4} Y_{4j}$
$\bar{Y}_{2.} = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_{2j}$
$\bar{Y}_{3.} = \frac{1}{n_3} \sum_{j=1}^{n_3} Y_{3j}$
$\bar{Y}_{1.} = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{1j}$
$\bar{Y}_{4.}$
$\bar{Y}_{2.}$
$\bar{Y}_{3.}$
$\bar{Y}_{1.}$

,

whereby the investigator imagines data coming from K different distributions such that the distributions are centered at the population treatment means $\mu_1, \dots, \mu_K$. Moreover, each distribution is Normal and all have the same variance, which is the variance σ_ε^2 of the error term in the cell means model. The sample treatment means $\bar{Y}_{1.}, \dots, \bar{Y}_{K.}$ are the investigator's guesses from the observed data at the unknown values of $\mu_1, \dots, \mu_K$.

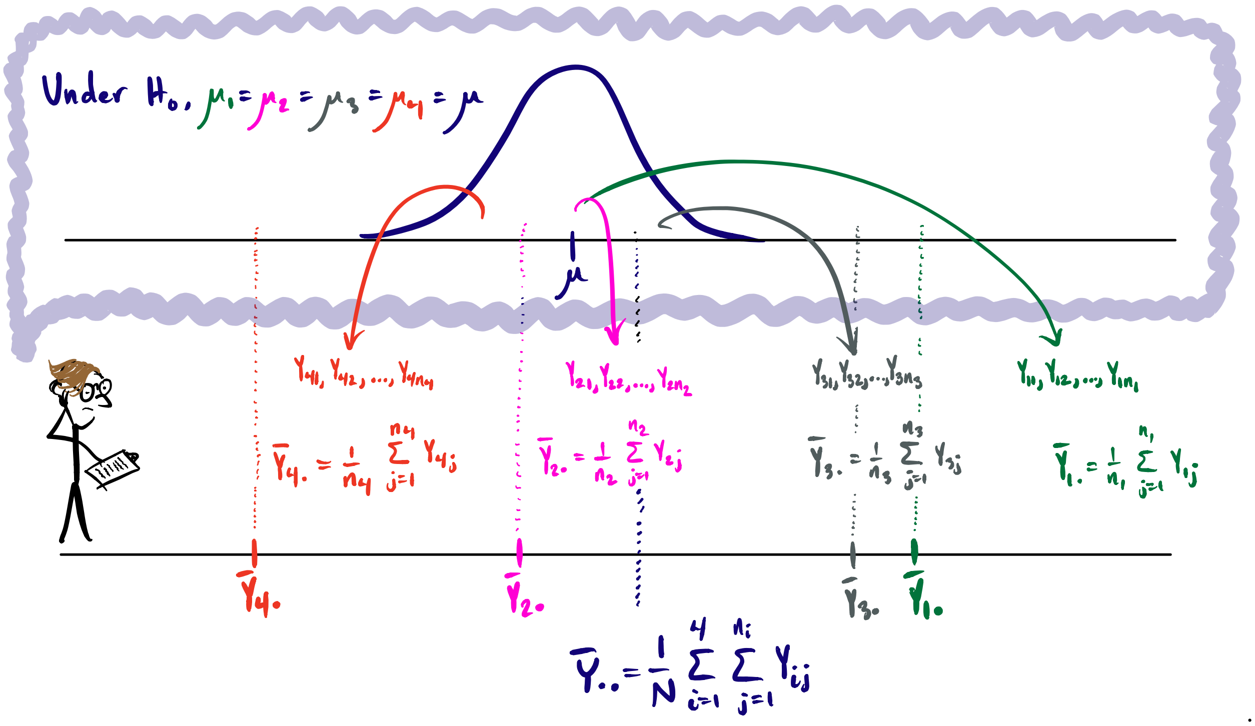
Hypothesis testing for the cell means model

In comparative experiments, the foremost research question is: Do/does any of the treatments affect the response? We can formulate this question in the following null and alternate hypotheses:

$$\begin{aligned} H_0\colon &\quad \mu_1 = \cdots = \mu_K \\ H_1\colon &\quad \mu_i \neq \mu_{i'} \text{ for some } i \neq i', \text{ i.e. not all treatment means are equal.} \end{aligned}$$

As we have previously done, we will test these hypotheses by computing a test statistic; once we have the test statistic we can compare it to a critical value or get a p-value from it in order to decide whether or not to reject H_0.

We may gather intuition for constructing a test statistic by picturing the cell means model under the null hypothess H_0: $\mu_1 = \cdots = \mu_K$. Returning to our example in which $K = 4$, the investigator now imagines all the data coming from a single distribution:


Under H_0, $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$
μ
$Y_{41}, Y_{42}, \ldots, Y_{4n_4}$
$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$
$Y_{31}, Y_{32}, \ldots, Y_{3n_3}$
$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$
$\bar{Y}_{4\cdot} = \frac{1}{n_4} \sum_{j=1}^{n_4} Y_{4j}$
$\bar{Y}_{2\cdot} = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_{2j}$
$\bar{Y}_{3\cdot} = \frac{1}{n_3} \sum_{j=1}^{n_3} Y_{3j}$
$\bar{Y}_{1\cdot} = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{1j}$
$\bar{Y}_{4\cdot}$
$\bar{Y}_{2\cdot}$
$\bar{Y}_{3\cdot}$
$\bar{Y}_{1\cdot}$
$\bar{Y}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{4} \sum_{j=1}^{n_i} Y_{ij}$

A new quantity which we may call the overall mean is introduced in the above picture. If it is assumed that all responses in all the treatment groups really come from a single distribution centered at a common mean μ, then a sensible estimator of the common mean is the mean of all the responses pooled together, which we write as

$$\bar{Y}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{n_i} Y_{ij}.$$

In Example with the steaks, we have

$$\bar{Y}_{\cdot\cdot} = \frac{1}{12}(7.66 + 6.89 + \ldots 3.66) = 5.9.$$

We now wish to construct a test statistic which gauges how much evidence the data carries against the null hypothesis. What would cast doubt on H_0? The more spread out the treatment means $\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{K\cdot}$ the more implausible H_0 becomes. So our test statistic should measure the spread of our treatment means $\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{K\cdot}$ and, under the null hypothesis, it should have a known probability distribution which allows us to look up a p-value.

The following result gives us some direction.

Result 1: Idealized test statistic assuming σ_ε known

nder the cell means model (1) under H_0: $\mu_1 = \cdots = \mu_K$, we have

$$\sum_{i=1}^{K} \left(\frac{\bar{Y}_{i.} - \bar{Y}_{..}}{\sigma_\varepsilon / \sqrt{n_i}} \right)^2 \sim \chi^2_{K-1}.$$

Recall that χ^2_{K-1} denotes the chi-squared distribution with degrees of freedom equal to $K - 1$.

The quantity in the above result measures how spread out the treatment means are, as it is the sum of squared deviations of the treatment means from the overall mean divided by $\sigma_\varepsilon/\sqrt{n}$. The larger this quantity is, the more implausible is H_0. Moreover, this quantity has a known distribution under H_0, which means we can find p-values; that is, we can find the probability, assuming that H_0 is true, of getting a larger value (carrying more evidence against H_0) of this quantity than the value we observed. The problem is that σ^2_ε is unknown, so we cannot compute this quantity from the data.

Define the estimator $\hat{\sigma}^2_\varepsilon$ of σ^2_ε as

$$\hat{\sigma}^2_\varepsilon = \frac{1}{N-K} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \hat{\varepsilon}^2_{ij},$$

where $\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_{i.}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, K$ are the deviations of the responses from their treatment means. These deviations are called *residuals*. The residuals $\hat{\varepsilon}_{ij}$ are like the sample version of the noise or error terms ε_{ij}, and the above formula uses them to estimate the variance σ^2_ε.

For Example with the steaks, the residuals are shown in the fourth column of the table below:

Steak	Packaging	log(# bact/cm^2)	$\bar{Y}_{i.}$	$\hat{\varepsilon}_{ij}$
1	Commercial	7.66	7.48	0.18
6	Commercial	6.98	7.48	−0.50
7	Commercial	7.80	7.48	0.32
12	Vacuum	5.26	5.50	−0.24
5	Vacuum	5.44	5.50	−0.06
3	Vacuum	5.80	5.50	0.30
10	Mixed Gas	7.41	7.26	0.15
9	Mixed Gas	7.33	7.26	0.07
2	Mixed Gas	7.04	7.26	−0.22
8	CO$_2$	3.51	3.36	0.15
4	CO$_2$	2.91	3.36	−0.45
11	CO$_2$	3.66	3.36	0.30

Using the residuals in the table, we would compute $\hat{\sigma}_\varepsilon^2$ as

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{12-4}\left[(0.18)^2 + (-0.50)^2 + \cdots + (0.30)^2\right] = 0.11585.$$

Then we have $\hat{\sigma}_\varepsilon = 0.34037$.

Now the next result presents a feasible test statistic ("feasible" meaning that we can actually compute it from the data) which has a distribution called an F distribution.

Result: Distribution of F-statistic under null hypothesis

Under the cell means model (1) under H_0: $\mu_1 = \cdots = \mu_K$, the quantity

$$F_{\text{test}} = \frac{1}{K-1}\sum_{i=1}^{K}\left(\frac{\bar{Y}_{i.} - \bar{Y}_{..}}{\hat{\sigma}_\varepsilon/\sqrt{n_i}}\right)^2$$

has the F-distribution with numerator degrees of freedom equal to $K-1$ and denominator degrees of freedom equal to $N-K$.

We will use this result to get a p-value for testing H_0: $\mu_1 = \cdots = \mu_K$, but first we introduce the F distributions.

For Example with the steaks, we have

$$\begin{aligned}F_{\text{test}} &= \frac{1}{4-1}\left[\left(\frac{7.48-5.9}{0.34037/\sqrt{3}}\right)^2 + \left(\frac{5.50-5.9}{0.34037/\sqrt{3}}\right)^2 + \left(\frac{7.26-5.9}{0.34037/\sqrt{3}}\right)^2 + \left(\frac{3.36-5.9}{0.34037/\sqrt{3}}\right)^2\right]\\ &= 94.58296.\end{aligned}$$

Is this large enough for us to reject H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$?

The F-distributions

An F-distribution arises when the ratio of two independent Chi-square random variables is taken such that each is divided by its degrees of freedom. Formally:

Result: The F random variable as a ratio of chi-squared random variables

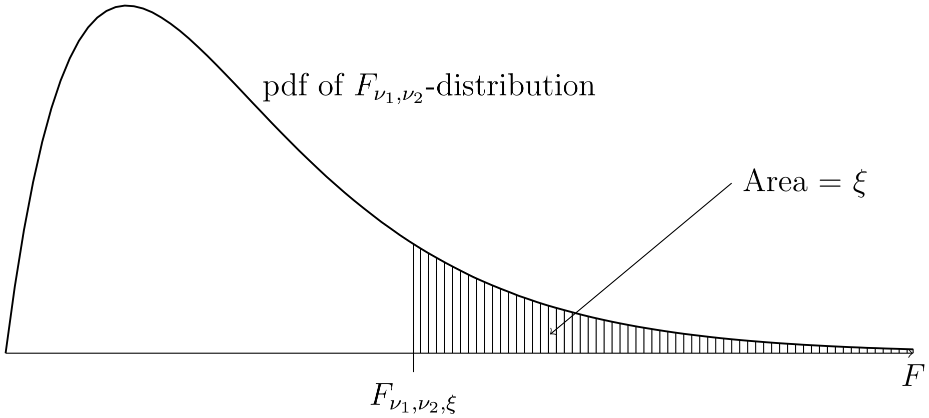
Let $W_1 \sim \chi^2_{\nu_1}$ and $W_2 \sim \chi^2_{\nu_2}$ be independent chi-squared random variables with degrees of freedom ν_1 and ν_2, respectively. Then

$$\frac{W_1/\nu_1}{W_2/\nu_2} \sim F_{\nu_1,\nu_2},$$

where F_{ν_1,ν_2} denotes the F-distribution with numerator degrees of freedom equal to ν_1 and denominator degrees of freedom equal to ν_2.

The F-distributions are right-skewed distributions and are indexed by two parameters called the numerator degrees of freedom and the denominator degrees of freedom which they inherit from the chi-squared distributions in the above definition. If a random variable F has the F-distribution with numerator degrees of freedom ν_1 and denominator degrees of freedom ν_2, we write $F \sim F_{\nu_1,\nu_2}$.

As we have defined upper quantiles for the standard Normal distribution, the t-distributions, and the chi-squared distributions, we define, for $0 < \xi < 1$, the value $F_{\nu_1,\nu_2,\xi}$ as the value such that $P(F > F_{\nu_1,\nu_2,\xi}) = \xi$, where $F \sim F_{\nu_1,\nu_2}$. The quantity $F_{\nu_1,\nu_2,\xi}$ thus admits the depiction



where the blue curve represents the probability density function of the F-distribution with numerator degrees of freedom ν_1 and denominator degrees of freedom ν_2.

Rejection rule for the equal-means hypothesis

Equipped with the F-distributions, we can establish a rejection rule for H_0: $\mu_1 = \cdots = \mu_K$. We will reject H_0 at significance level α if

$$F_{\text{test}} = \frac{1}{K-1}\sum_{i=1}^{K}\left(\frac{\bar{Y}_{i.} - \bar{Y}_{..}}{\hat{\sigma}_\varepsilon/\sqrt{n_i}}\right)^2 > F_{K-1,N-K,\alpha}.$$

That is, we can pull the critical value $F_{K-1,N-K,\alpha}$ from the F distribution with numerator degrees of freedom $K-1$ and denominator degrees of freedom $N-K$ and compare our test statistic F_{test} to it.

We can also simply compute the p-value associated with the value of the test statistic F_{test}. The p-value is equal to the probability

$$P(F > F_{\text{test}}), \quad \text{where } F \sim F_{K-1,N-K},$$

which is the area under the pdf of the $F_{K-1,N-K}$ distribution to the right of F_{test}.

In Example with the steaks we have $F_{\text{test}} = 94.58$ (after rounding). Since there are $K = 4$ treatments and $N = 12$ experimental units, we compare the value of F_{test} to the upper α quantile of the F-distribution with numerator degrees of freedom $4-1 = 3$ and denominator degrees of freedom $12-4 = 8$. For $\alpha = 0.01$ we have $F_{3,8,0.01} = 7.59$ (see the F-table on pg. 826 of the textbook). Since $F_{\text{test}} = 94.58 > F_{3,8,0.01} = 7.59$ we would reject the null hypothesis at sigificance level 0.01 and conclude that not all the means are equal.

The p-value is the probability

$$P(F > 94.58), \quad \text{where } F \sim F_{3,8}.$$

We can get this probability using the R function `pf()`, which returns F-distribution probabilities such that

$$\texttt{pf(q,df1,df2)} = P(F < \texttt{q}), \quad \text{where } F \sim F_{\texttt{df1},\texttt{df2}}.$$

Thus the p-value is given by

$$\texttt{1 - pf(94.584,3,8)} = 1.375902 \times 10^{-6}.$$

If the assumptions of the cell means model are satisfied for the steaks example, then the small p-value indicates very strong evidence against the null hypothesis of equal treatment means. We would conclude at any significance level α greater than 1.376128×10^{-6} that the packaging makes a difference in the mean number of bacteria that grow on the surface of refrigerated steaks.

Analysis of variance (ANOVA)

The analysis of variance or ANOVA table decomposes the so-called total variation in the responses Y_{ij}, $j = 1, \dots, n_i$, $i = 1, \dots, K$ into two parts:

1. *Between-treatment variation*: variability in the responses due to the treatment effects.
2. *Within-treatment variation*: variability in the responses due to differences among the experimental units.

We will express the total variability in the response values through the quantity

$$\mathrm{SS}_{\mathrm{Total}} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2,$$

which we will call the *total sum of squares*. This quantity is the sum of squared deviations of the responses from the overall mean $\bar{Y}_{..}$. The ANOVA approach decomposes this quantity into the *treatment sum of squares*

$$\mathrm{SS}_{\mathrm{Treatment}} = \sum_{i=1}^{K} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

and the *error sum of squares*

$$\mathrm{SS}_{\mathrm{Error}} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2.$$

Verily

$$\underbrace{\mathrm{SS}_{\mathrm{Total}}}_{\text{Total}} = \underbrace{\mathrm{SS}_{\mathrm{Treatment}}}_{\text{Between}} + \underbrace{\mathrm{SS}_{\mathrm{Error}}}_{\text{Within}},$$

where $\mathrm{SS}_{\mathrm{Treatment}}$ represents between-treatment variation and $\mathrm{SS}_{\mathrm{Error}}$ represents within-treatment variation. For these quantities, we have the following:

Result: Chi-squared distributions of scaled sums of squares

Under the cell means model (1) under H_0: $\mu_1 = \cdots = \mu_K$, we have

$$\begin{aligned} \mathrm{SS}_{\mathrm{Treatment}}/\sigma_\varepsilon^2 &\sim \chi^2_{K-1} \\ \mathrm{SS}_{\mathrm{Error}}/\sigma_\varepsilon^2 &\sim \chi^2_{N-K}, \end{aligned}$$

where $N = n_1 + \cdots + n_K$.

We find that we can construct our test statistic F_{test} in terms of these quantities; that is

$$F_{\text{test}} = \frac{1}{K-1}\sum_{i=1}^{K}\left(\frac{\bar{Y}_{i.} - \bar{Y}_{..}}{\hat{\sigma}_{\varepsilon}/\sqrt{n_i}}\right)^2 = \frac{\text{SS}_{\text{Treatment}}\,/(K-1)}{\text{SS}_{\text{Error}}\,/(N-K)},$$

such that the test statistic F_{test} is the ratio of between-to-within-treatment variability. Let us give names to the quantities in the numerator and denominator of the last expression; that is, let

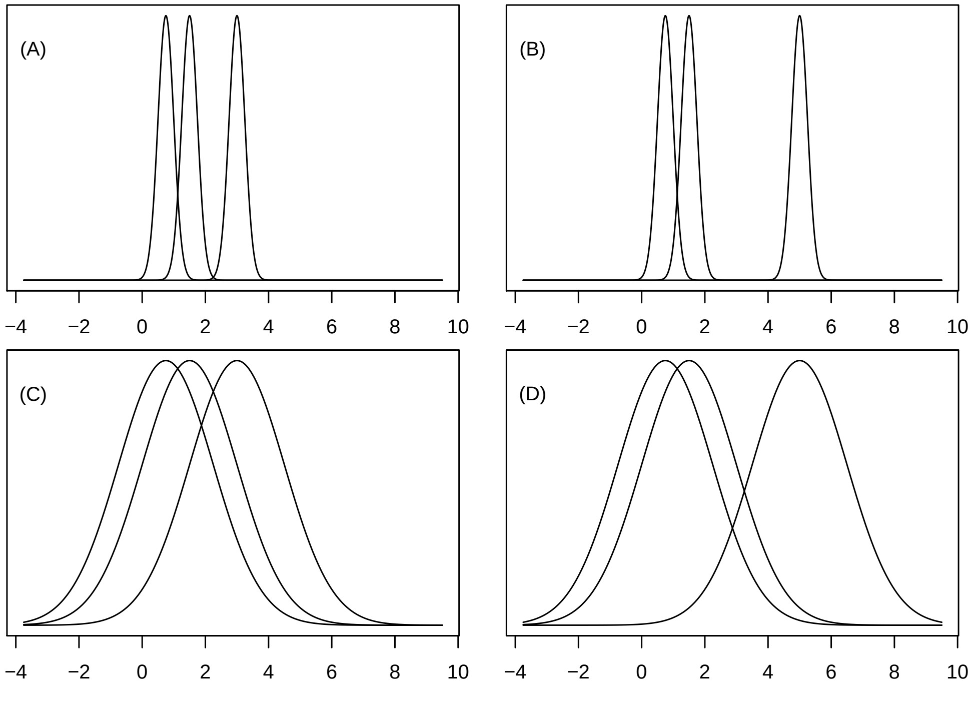
$$\begin{aligned}\text{MS}_{\text{Treatment}} &= \text{SS}_{\text{Treatment}}\,/(K-1)\\ \text{MS}_{\text{Error}} &= \text{SS}_{\text{Error}}\,/(N-K),\end{aligned}$$

and let us call these, respectively, the *treatment mean square* and the *mean squared error*. In the end we may write

$$F_{\text{test}} = \frac{\text{MS}_{\text{Treatment}}}{\text{MS}_{\text{Error}}} \quad \left(= \frac{\text{Between-treatment variation}}{\text{Within-treatment variation}}\right).$$

If the treatment means are very spread out relative to the spread of the responses in each treatment group, the test statistic F_{test} will be large, and its p-value will be small. If the treatment means are close together relative to the spread of the responses in each treatment group, the test statistic F_{test} will be small, and its p-value will be large. Consider the following illustration.

Example. Consider the three probability density functions depicted in each panel of the following figure as the distributions of the responses in a cell means or one-way ANOVA model with $K = 3$ treatments.



Suppose a randomized experiment is conducted for testing the hypotheses H_0: $\mu_1 = \mu_2 = \mu_3$ versus H_1: not all means are equal.

1. Of the settings depicted in the four panels, from which would we expect the largest value of the test statistic F?
   The setting in panel (B) exhibits large between-treatment variation and small within-treatment variation, so we would expect from it the largest value of the test statistic F_{test}.
2. Of the settings depicted in the four panels, from which would we expect the smallest value of the test statistic F?
   The setting in panel (C) exhibits small between-treatment variation and large within-treatment variation, so we would expect it to produce the smallest value of the test statistic F_{test}.
3. Which two settings are likely to result in greater measures of between-treatment variation $\text{MS}_{\text{Treatment}}$ than the other two?
   The settings in panels (B) and (D) exhibit larger between-treatment variation than the other two panels, so we would expect them to produce larger values of $MS_{\text{Treatment}}$.
4. Which two settings are likely to result in greater measures of within-treatment variation MS_{Error} than the other two?
   The settings in panels (C) and (D) exhibit larger within-treatment variation than the other two panels, so we would expect them to produce larger values of MS_{Error}.

The ANOVA table

But what about the table? The ANOVA table is a conventional format in which to present the several quantities introduced in this section. The ANOVA table for the cell means model looks like this:

Source	Sum of Squares	df	Mean Squares	F	p-value
Treatment	$\text{SS}_{\text{Treatment}}$	$K-1$	$\text{MS}_{\text{Treatment}}$	F_{test}	$P(F > F_{\text{test}})$
Error	SS_{Error}	$N-K$	MS_{Error}		where $F \sim F_{K-1,N-K}$
Total	SS_{Total}	$N-1$			

For the steaks example, the ANOVA table is

Source	Sum of Squares	df	Mean Squares	F	p-value
Treatment	32.873	3	10.9576	94.584	1.375902×10^{-6}
Error	0.927	8	0.1159		
Total	33.800	11			

This can be obtained from R from the commands below

```
# read in the data and format it for ANOVA:
bacteria <- c(7.66,6.98,7.80,5.26,5.44,5.80,7.41,7.33,7.04,3.51,2.91,3.66)
packaging <- c( rep("Commercial",3),
                rep("Vacuum",3),
                rep("Mixed Gas",3),
                rep("C02",3))
packaging <- as.factor(packaging)

# estimate the cell means model with lm() function and retrieve ANOVA table:
model <- lm(bacteria ~ packaging)
anova(model)
```

The output looks like the following (Note that R arranges the columns in the ANOVA table somewhat differently and does not provide the bottom row.):

```
Analysis of Variance Table

Response: bacteria
          Df Sum Sq Mean Sq F value    Pr(>F)
packaging  3 32.873 10.9576  94.584 1.376e-06 ***
Residuals  8  0.927  0.1159
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We will see that the ANOVA tables appear in other contexts as well; they are not intended solely for the cell-means model.

Checking the assumptions of the cell means model

Implicit in the definition of the cell means model

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma_\varepsilon^2), \quad j = 1, \dots, n_i, \quad i = 1, \dots, K$$

are some important assumptions about the data-generating mechanism. We will call them Assumptions (A.1), (A.2), and (A.3):

(A.1) The responses are Normally distributed around the treatment means.

To check: Look at a Normal QQ plot of the residuals

$$\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_{i.}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, K.$$

The residuals are the deviations of the responses around the sample treatment means, so they are the sample version of the error terms ε_{ij}, which are assumed to be Normal. The residuals should therefore exhibit Normality in the Normal QQ plot.

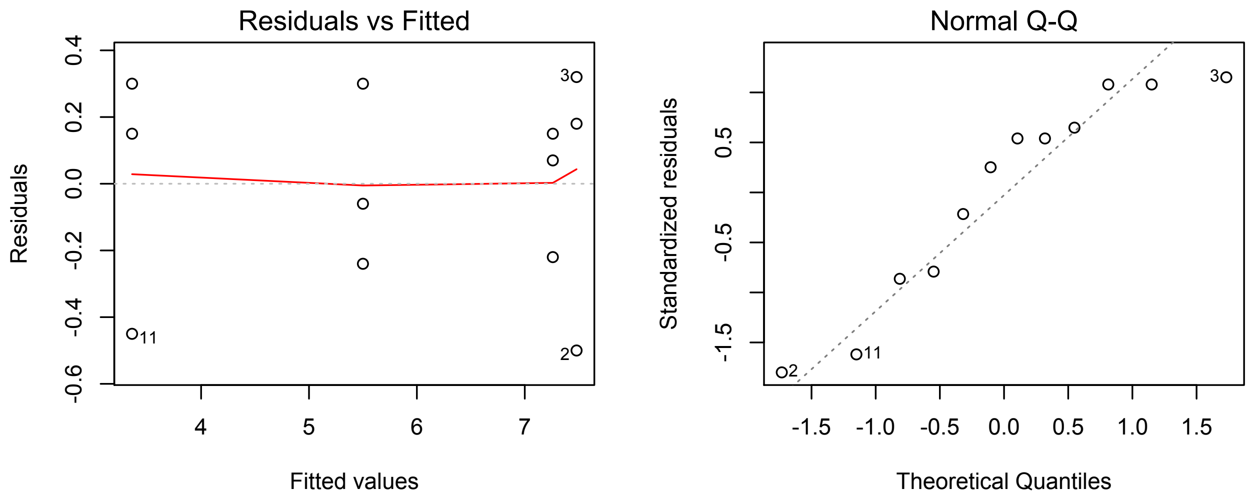
(A.2) The variance of the responses is the same in each treatment group.

To check: Look at the *residuals versus fitted values* plot. In this plot, the residuals for each treatment group are plotted on the vertical axis against the treatment means such that the spread of the residuals across the treatment groups can be compared. The spreads of the residuals should be the same across the treatment groups.

(A.3) The responses are independent from each other.

Cannot check: This assumption cannot be checked from the data, but the random assignment of the experimental units to the treatment groups and proper control of experimental conditions should ensure that the responses are independent from each other.

Continuing with the steaks example, the command `plot(lm(bacteria ~ packaging))` causes R to produce a series of plots, among which are the Normal QQ plot and the residuals versus fitted values plot:



It is difficult to see whether the spread of the residuals is the same in all four treatment groups, as there are only three experimental units in each group, but the spreads are not drastically different, so we may assume that (A.2) holds. The residuals look to be Normally distributed from the Normal QQ plot, so we may assume that (A.1) holds.

References

[1] R. O. Kuehl. *Design of Experiments: Statistical Principles of Research Design and Analysis*. Duxbury/Thomson Learning, 2000. Google-Books-ID: mIV2QgAACAAJ.