Introduction to Analysis of Variance

Eli Gurarie

QERM 598 - Lecture 3

University of Washington - Seattle

January 23, 2008

Preface: Some Necessary Math Facts

- if $X_1, X_2, ...X_n$ are iid rv's with distribution $\mathrm{N}\left\{\mu, \sigma^2\right\}$ then:

$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\mu} \sim \mathrm{N}\left\{\mu, \frac{\sigma^2}{n}\right\} \tag{1}$$

$$S^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \hat{\sigma}^2 \tag{2}$$

$$\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \text{Chi-squared}\{n-1\} \tag{3}$$

- if $Y_1, Y_2, ...Y_n$ and $Z_1, Z_2, ...Z_m$ are iid rv's with distribution $\mathrm{N}\{0,1\}$ then:

$$\frac{\frac{1}{m}\sum_{i=1}^{m} Z_i^2}{\frac{1}{n}\sum_{i=1}^{n} Y_i^2} \quad \sim \quad \mathrm{F}\{m,n\} \tag{4}$$

- **Cochran's Theorem**: If Z_i are iid $\mathrm{N}\{0,1\}$ for $i = 1, 2, ..., \nu$ AND

$$\sum_{i=1}^{\nu} Z_i^2 = Q_1 + Q_2 + ... + Q_s \tag{5}$$

where Q_i is the sum of ν_i squared random variables AND $\nu = \nu_1 + \nu_2 + ... + \nu_s$, THEN, $Q_1, Q_2, ..., Q_s$ are independent chi-squared random variables with $\nu_1, \nu_2, ..., \nu_s$ degrees of freedom.

Historical roots of ANOVA

Sir Ronald Aylmer Fisher (1890-1962) was one of the greatest statisticians and population geneticists of the 20th century*, the main developer of ANOVA, the namesake of the F-distribution, and source of many many other contributions. Since Fisher's main interest was genetics, he was interested in relating differences in *phenotype* to differences in *genotype*. The presence of different *alleles* (versions of a gene) are a discrete factor which are often expressed in continuous phenotypes, such as height, weight, or pigment. Out of this problem arose the extremely useful and versatile family of models known as ANOVA.

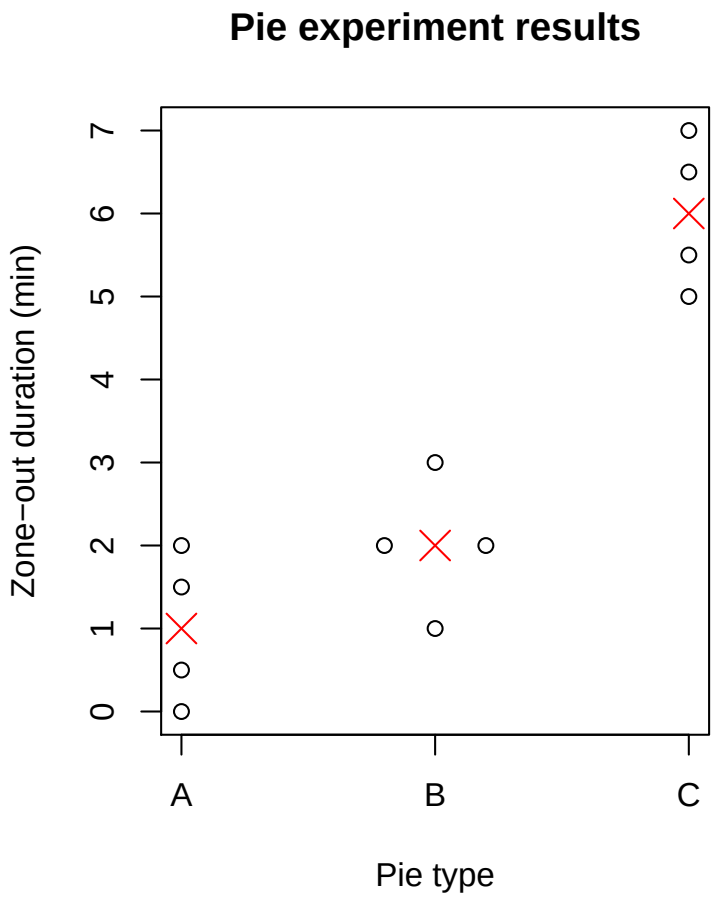
*- also, a big advocate of human eugenics and by many accounts a difficult person to deal with. I guess it takes all kinds.

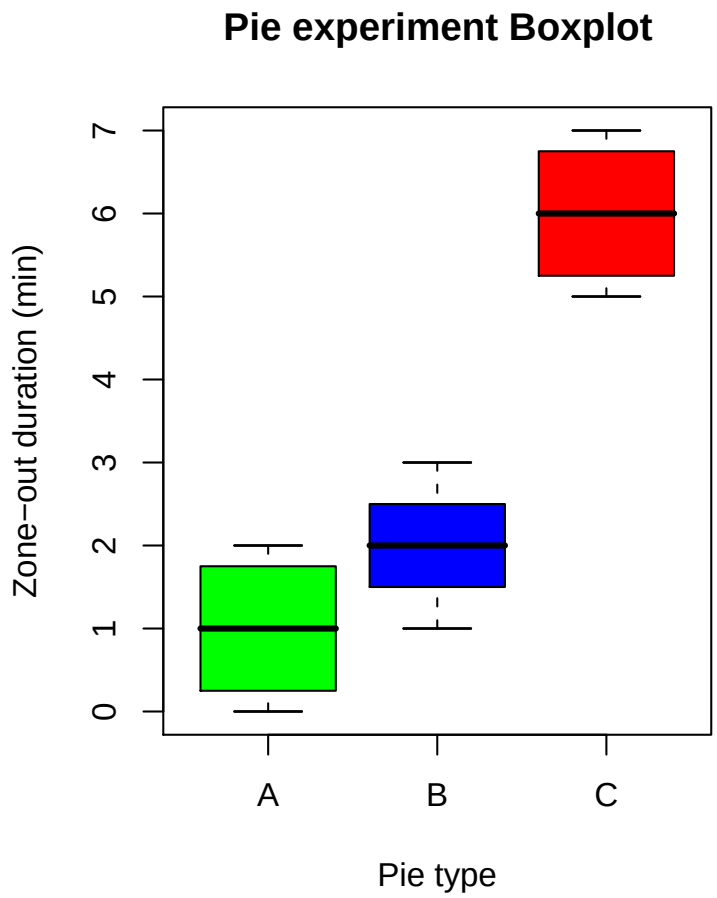
The Great Pie Zone-out Experiment

One Wednesday after the weekly QERM soup, an experiment was performed to test the effects of different desserts on student concentration. The twelve students were divided into three groups of four, each of what was to consume in its entirety an Apple pie, a Blueberry pie, and a Cherry pie. Later, from 1:30-3:00 all twelve students attended a statistics department seminar. All but one of the students zoned out at least once during the seminar, and the total zone-outs duration (ZOD) in minutes was carefully recorded by the experimenter. The results (in minutes) are tabulated below:

Treatment	**ZOD (min)**				**totals**	**means ($\bar{x}_{i\cdot}$)**
Apple Pie	0	2	0.5	1.5	4	1.0
Blueberry Pie	1	2	3	2	8	2.0
Cherry Pie	7	5.5	6.5	5	24	6.0
totals					**36**	**3.0**

Visualization





Formulating a hypothesis

- Research question:
  Do different kinds of dessert have different effects on the concentration of QERM students?
- Null hypothesis in words:
  Different pie treatments result in essentially similar ZOD's
- Alternative hypothesis in words:
  Different pie treatments result in different ZOD's
- Null hypothesis in math:
  $\mu_A = \mu_B = \mu_C$
- Alternative hypothesis in math:
  $\mu_A \neq \mu_B$ OR $\mu_A \neq \mu_C$ OR $\mu_B \neq \mu_C$

Formulating a hypothesis

- Research question:
  Do different kinds of dessert have different effects on the concentration of QERM students?
- Null hypothesis in words:
  Different pie treatments result in essentially similar ZOD's
- Alternative hypothesis in words:
  Different pie treatments result in different ZOD's
- Null hypothesis in math:
  $\mu_A = \mu_B = \mu_C$
- Alternative hypothesis in math:
  $\mu_A \neq \mu_B$ OR $\mu_A \neq \mu_C$ OR $\mu_B \neq \mu_C$

Formulating a hypothesis

- Research question:
  Do different kinds of dessert have different effects on the concentration of QERM students?
- Null hypothesis in words:
  Different pie treatments result in essentially similar ZOD's
- Alternative hypothesis in words:
  Different pie treatments result in different ZOD's
- Null hypothesis in math:
  $\mu_A = \mu_B = \mu_C$
- Alternative hypothesis in math:
  $\mu_A \neq \mu_B$ OR $\mu_A \neq \mu_C$ OR $\mu_B \neq \mu_C$

Formulating a hypothesis

- Research question:
  Do different kinds of dessert have different effects on the concentration of QERM students?
- Null hypothesis in words:
  Different pie treatments result in essentially similar ZOD's
- Alternative hypothesis in words:
  Different pie treatments result in different ZOD's
- Null hypothesis in math:
  $\mu_A = \mu_B = \mu_C$
- Alternative hypothesis in math:
  $\mu_A \neq \mu_B$ OR $\mu_A \neq \mu_C$ OR $\mu_B \neq \mu_C$

Formulating a hypothesis

- Research question:
  Do different kinds of dessert have different effects on the concentration of QERM students?
- Null hypothesis in words:
  Different pie treatments result in essentially similar ZOD's
- Alternative hypothesis in words:
  Different pie treatments result in different ZOD's
- Null hypothesis in math:
  $\mu_A = \mu_B = \mu_C$
- Alternative hypothesis in math:
  $\mu_A \neq \mu_B$ OR $\mu_A \neq \mu_C$ OR $\mu_B \neq \mu_C$

Comments on models and hypotheses

- So far, we've formulated scientific questions in terms of hypotheses and hypothesis tests, consistent with *t*-tests, randomization, Mann-Whitney, etc. to compare samples.
- When confronted with more complicated systems or datasets, hypothesis-testing is a little narrow. It is more enlightening to think in terms of *model assessment*. We typically propose several possible *statistical models* and assess which has greater explanatory power given the quality of the data. In practice, you will be looking at the results of large ANOVA tables which can contain within them many implicit hypotheses and are all essentially assessed simultaneously as we select the 'best' or most parsimonious model. The hypothesis test is best thought of as a *tool* in the *model selection process*.
- This distinction is reflected in the nomenclature. Even a very simple design like the pie experiment, where the analysis is just one step more complicated than a two-sample *t*-test, we use an *ANALYSIS* of variance, whereas the *t*-test is 'merely' a *TEST*.

Comments on models and hypotheses

- So far, we've formulated scientific questions in terms of hypotheses and hypothesis tests, consistent with *t*-tests, randomization, Mann-Whitney, etc. to compare samples.
- When confronted with more complicated systems or datasets, hypothesis-testing is a little narrow. It is more enlightening to think in terms of *model assessment*. We typically propose several possible *statistical models* and assess which has greater explanatory power given the quality of the data. In practice, you will be looking at the results of large ANOVA tables which can contain within them many implicit hypotheses and are all essentially assessed simultaneously as we select the 'best' or most parsimonious model. The hypothesis test is best thought of as a *tool* in the *model selection process*.
- This distinction is reflected in the nomenclature. Even a very simple design like the pie experiment, where the analysis is just one step more complicated than a two-sample *t*-test, we use an *ANALYSIS* of variance, whereas the *t*-test is 'merely' a *TEST*.

Comments on models and hypotheses

- So far, we've formulated scientific questions in terms of hypotheses and hypothesis tests, consistent with *t*-tests, randomization, Mann-Whitney, etc. to compare samples.
- When confronted with more complicated systems or datasets, hypothesis-testing is a little narrow. It is more enlightening to think in terms of *model assessment*. We typically propose several possible *statistical models* and assess which has greater explanatory power given the quality of the data. In practice, you will be looking at the results of large ANOVA tables which can contain within them many implicit hypotheses and are all essentially assessed simultaneously as we select the 'best' or most parsimonious model. The hypothesis test is best thought of as a *tool* in the *model selection process*.
- This distinction is reflected in the nomenclature. Even a very simple design like the pie experiment, where the analysis is just one step more complicated than a two-sample *t*-test, we use an *ANALYSIS* of variance, whereas the *t*-test is 'merely' a *TEST*.

Formulating a statistical model

- Model 1 - Single mean: $X_{ij} = \mu + \epsilon_{ij}$
- Model 2 - Unique means: $X_{ij} = \mu_i + \epsilon_{ij}$

Where:

- X_{ij} uniquely identifies an individual measurement.
- $i \in \{1, 2...a\}$ indexes the treatment. Here, $a = 3$, representing pies A, B and C.
- $j \in \{1, 2...n\}$ indexes the individual measurement within each treatment group. Here, $n = 4$, and the total number of samples $N = an$:
- μ is the true grand mean;
- μ_i is a true group mean within each treatment group;
- ϵ_{ij} is a random individual error term

Very, very important assumption: ϵ_{ij}'s are i.i.d. $\mathrm{N}\left\{0, \sigma^2\right\}$.

Formulating a statistical model

- Model 1 - Single mean: $X_{ij} = \mu + \epsilon_{ij}$
- Model 2 - Unique means: $X_{ij} = \mu_i + \epsilon_{ij}$

Where:

- X_{ij} uniquely identifies an individual measurement.
- $i \in \{1, 2...a\}$ indexes the treatment. Here, $a = 3$, representing pies A, B and C.
- $j \in \{1, 2...n\}$ indexes the individual measurement within each treatment group. Here, $n = 4$, and the total number of samples $N = an$:
- μ is the true grand mean;
- μ_i is a true group mean within each treatment group;
- ϵ_{ij} is a random individual error term

Very, very important assumption: ϵ_{ij}'s are i.i.d. $\mathrm{N}\left\{0, \sigma^2\right\}$.

Formulating a statistical model

- Model 1 - Single mean: $X_{ij} = \mu + \epsilon_{ij}$
- Model 2 - Unique means: $X_{ij} = \mu_i + \epsilon_{ij}$

Where:

- X_{ij} uniquely identifies an individual measurement.
- $i \in \{1, 2...a\}$ indexes the treatment. Here, $a = 3$, representing pies A, B and C.
- $j \in \{1, 2...n\}$ indexes the individual measurement within each treatment group. Here, $n = 4$, and the total number of samples $N = an$:
- μ is the true grand mean;
- μ_i is a true group mean within each treatment group;
- ϵ_{ij} is a random individual error term

Very, very important assumption: ϵ_{ij}'s are i.i.d. $\mathrm{N}\left\{0, \sigma^2\right\}$.

Comparing Variances

The goal of ANOVA is to compare *within group* variances ($S_{i\cdot}^2$) and *between group* variances ($S_{\cdot\cdot}^2$). If the within-group variance is somehow smaller than the between-group variance, then the treatment might have some explanatory power, i.e. a certain amount of the variance is accounted for by the treatment effect.
Consider our data:

Tr.	**ZOD (min)**				**means** ($\bar{x}_{i\cdot}$)	**variances** ($S_{i\cdot}^2$)
A	0	2	0.5	1.5	1.0	0.8333
B	1	2	3	2	2.0	0.6667
C	7	5.5	6.5	5	6.0	0.8333
					$\bar{v}_{\cdot\cdot} = 3.0$	$S_{\cdot\cdot}^2 = 5.7272$

The values for $S_{i\cdot}^2$ certainly *look* smaller than $S_{\cdot\cdot}^2$. But how do we perform rigorous inference?

Several Sums of Squares

The *sum of squares* is a total measure of variability. Consider the *Total* sum of squares *:

$$SS = \sum_{i=1}^{a} \sum_{j=1}^{n} (x_{ij} - \bar{x}_{..})^2 \tag{6}$$

* - note that $SS/(N-1)$ is an unbiased estimate of the total sample variance σ_p^2 of the data.

This sum can be decomposed into:

$$\begin{aligned} SS &= n \sum_{i=1}^{a} (x_{i.} - \bar{x}_{..})^2 + \sum_{i=1}^{a} \sum_{j=1}^{n} (x_{ij} - \bar{x}_{i.})^2 && (7) \\ &= SS_{treatment} + SS_{error} && (8) \end{aligned}$$

SS_{error} is the Error sum of squares, i.e. the sum of all the little deviations from each of the local little means, while $SS_{treatment}$ is the Treatment sum of squares, i.e. the sum of the differences of the little means from the grand mean, weighted according to the number of measurements within each group (n).

Several Sums of Squares

The *sum of squares* is a total measure of variability. Consider the *Total* sum of squares *:

$$SS = \sum_{i=1}^{a} \sum_{j=1}^{n} (x_{ij} - \bar{x}_{..})^2 \tag{6}$$

* - note that $SS/(N-1)$ is an unbiased estimate of the total sample variance σ_p^2 of the data.

This sum can be decomposed into:

$$\begin{aligned} SS &= n \sum_{i=1}^{a} (x_{i.} - \bar{x}_{..})^2 + \sum_{i=1}^{a} \sum_{j=1}^{n} (x_{ij} - \bar{x}_{i.})^2 && (7) \\ &= SS_{treatment} + SS_{error} && (8) \end{aligned}$$

SS_{error} is the Error sum of squares, i.e. the sum of all the little deviations from each of the local little means, while $SS_{treatment}$ is the Treatment sum of squares, i.e. the sum of the differences of the little means from the grand mean, weighted according to the number of measurements within each group (n).

Expectations of Sums of Squares 1

Under hypothesis that variances are iid:

$$\mathrm{E}\left[SS_{error}\right] = \mathrm{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\cdot})^2\right] = \sum_{i=1}^{a}(n-1)\sigma_i^2 = a(n-1)\sigma^2$$

Similarly,

$$\mathrm{E}\left[SS_{treatment}\right] = (a-1)\,\sigma^2 \tag{9}$$

Thus:

- $SS_{error}/(N-a)$ (called Mean Squared Error - MSE) is unbiased estimate of σ^2
- $SS_{treatment}/(a-1)$ (called Mean Squared Treatment - MST) is *ALSO* an unbiased estimate of σ^2.

Expectations of Sums of Squares 1

Under hypothesis that variances are iid:

$$\mathrm{E}\left[SS_{error}\right] = \mathrm{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\cdot})^2\right] = \sum_{i=1}^{a}(n-1)\sigma_i^2 = a(n-1)\sigma^2$$

Similarly,

$$\mathrm{E}\left[SS_{treatment}\right] = (a-1)\,\sigma^2 \tag{9}$$

Thus:

- $SS_{error}/(N-a)$ (called Mean Squared Error - MSE) is unbiased estimate of σ^2
- $SS_{treatment}/(a-1)$ (called Mean Squared Treatment - MST) is *ALSO* an unbiased estimate of σ^2.

Getting close

Anyways, we know:

$$\mathrm{E}\left[SS\right] = \mathrm{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{..})\right] = (n-1)\,\sigma_p^2 \tag{10}$$

Thus, applying Fact 3:

$$\frac{SS}{\sigma_p^2} \sim \text{Chi-squared}\{N-1\} \tag{11}$$

Under the null assumption (NO treatment effect) $\sigma_p^2 = \sigma^2$. Thus, since:

$$\frac{SS}{\sigma_p^2} = \frac{SS_{error}}{\sigma^2} + \frac{SS_{treatment}}{\sigma^2} \tag{12}$$

then (by Cochran's theorem) SS_{error}/σ^2 and $SS_{treatment}/\sigma^2$ must be independent Chi-squared random variables, with $N-a$ and $a-1$ degrees of freedom respectively.

Meet William Gemmell Cochran (1909-1980)

Cochran was a good statistician and, later, a great statistics department administrator. I was hoping to find something out-of-the-ordinary in his biography, but absolutely nothing came up. Really, I put his picture into these notes to break up the stream of equations.

Finally ...

we obtain the test statistic:

$$F_0 = \frac{SS_{treatment}/(a-1)}{SS_{error}/(N-a)} = \frac{MS_{treatment}}{MS_{error}} \tag{13}$$

Under H_0:

- MS_{error} and the $MS_{treatment}$ are unbiased estimators of σ^2
- $F_0 \sim \mathrm{F}\{a-1, N-a\}$ (by Fact 4)

Under H_1, $MS_{treatment} > MS_{error}$, and we can reject the null hypothesis based on comparing the F_0test statistic to the null distribution: $\mathrm{F}\{a-1, N-a\}$.

ANOVA table

Typically, we construct a table to summarize our analysis:

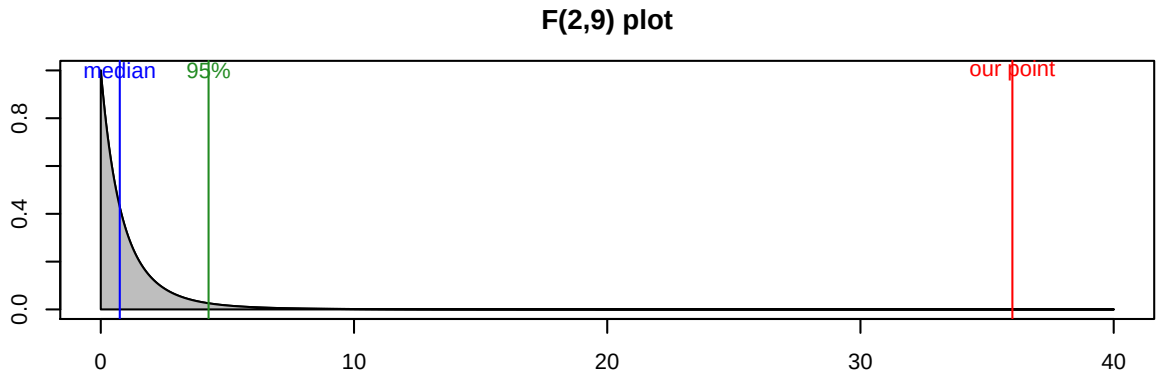
Source of Variation	**Sum of Squares**	**Degrees of freedom**	**Mean Square**	F_0
Treatment	SS_{treat}	$a-1$	MS_{treat}	$\frac{MS_{treat}}{MS_{error}}$
Residual	SS_{error}	$N-a$	MS_{error}	
Total	SS	$N-1$		

Treatment refers to variation explained by differences *between* group means,
Residuals refers to differences *within* group means.

Pie analysis

Pie experiment ANOVA table:

Source	SS	df	MS	F_0	p-value
Pie	56	2	28	36	5.081e-05
Residuals	7	9	0.778		
Total	63	11			



Under the null hypothesis, the F_0 statistic will be not significantly different from 1. Ours clearly looks extreme.
What is the probability of an even more extreme F_0 value emerging from a true null hypothesis?

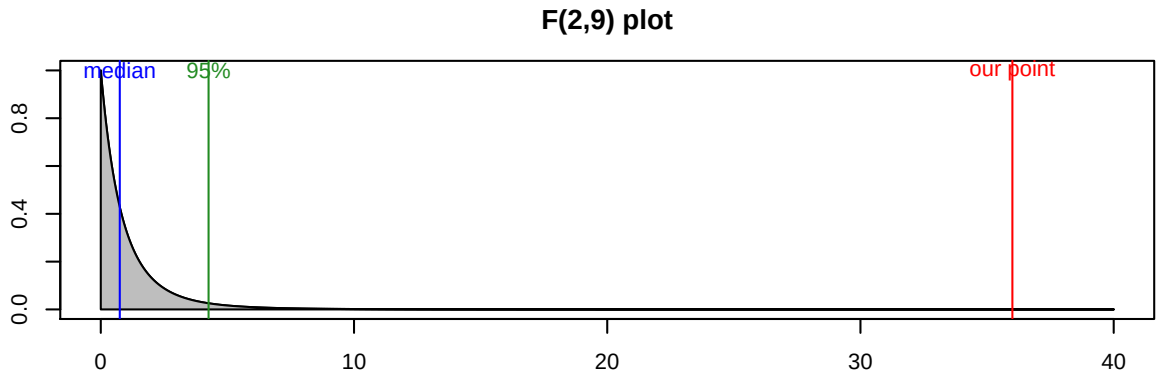
$$\Pr[F_0 > F_{2,9}] = 5.081 \times 10^{-05} \tag{14}$$

This is the p-value, or probability of Type I error. It is clearly very very small .... so we reject the null-hypothesis with great confidence.

Pie analysis

Pie experiment ANOVA table:

Source	SS	df	MS	F_0	p-value
Pie	56	2	28	36	5.081e-05
Residuals	7	9	0.778		
Total	63	11			



Under the null hypothesis, the F_0 statistic will be not significantly different from 1. Ours clearly looks extreme.

What is the probability of an even more extreme F_0 value emerging from a true null hypothesis?

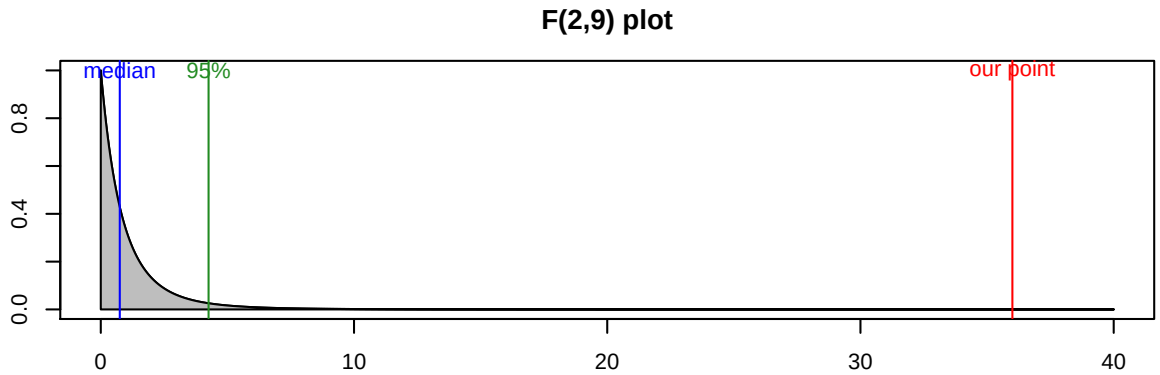
$$\Pr[F_0 > F_{2,9}] = 5.081 \times 10^{-05} \tag{14}$$

This is the p-value, or probability of Type I error. It is clearly very very small .... so we reject the null-hypothesis with great confidence.

Pie analysis

Pie experiment ANOVA table:

Source	SS	df	MS	F_0	p-value
Pie	56	2	28	36	5.081e-05
Residuals	7	9	0.778		
Total	63	11			



Under the null hypothesis, the F_0 statistic will be not significantly different from 1. Ours clearly looks extreme.
What is the probability of an even more extreme F_0 value emerging from a true null hypothesis?

$$\Pr[F_0 > F_{2,9}] = 5.081 \times 10^{-05} \tag{14}$$

This is the p-value, or probability of Type I error. It is clearly very very small .... so we reject the null-hypothesis with great confidence.

Model Specification

Recall our 2 models:

- Model 1 - Single mean: $X_{ij} = \mu + \epsilon_{ij}$
- Model 2 - Unique means: $X_{ij} = \mu_i + \epsilon_{ij}$

The ANOVA helped us pick out the most informative model (Model 2). It basically told us that accounting for the groups made our estimate for σ significantly smaller than not accounting for the groups would have.
Now that we have selected a model, we can *specify* it. In the model we've selected, there are 4 parameters: μ_1, μ_2, μ_3 and σ^2. The estimates for these parameters are:

parameter	estimate	value
μ_1 (Apple pie)	$\bar{x}_{1\cdot}$	1
μ_2 (Blueberry pie)	$\bar{x}_{2\cdot}$	2
μ_3 (Cherry pie)	$\bar{x}_{3\cdot}$	6
σ^2	MS_{error}	0.778

Hypotheses vs. Models

- Strictly speaking, the hypothesis test lets us say that: *There is at least one pair of means in our experiment that is not equal.* This is a relatively crude result, but it can be stated with great certainty.
- In contrast, the model we have selected lets us say that: *Given the data collected, we can predict that the mean effects of Apple, Blueberry and Cherry pie dosage on QERM students are about 1, 2 and 6 minutes of zoning out with some roughly normally distributed variability with variance around 0.8.*
- This second statement is not strictly speaking true. Like all models, it is a reduction and simplification of reality. However, given the information that we have, it is probably the best description of reality. The hypothesis test was an aid in selecting this model.

Hypotheses vs. Models

- Strictly speaking, the hypothesis test lets us say that: *There is at least one pair of means in our experiment that is not equal.* This is a relatively crude result, but it can be stated with great certainty.
- In contrast, the model we have selected lets us say that: *Given the data collected, we can predict that the mean effects of Apple, Blueberry and Cherry pie dosage on QERM students are about 1, 2 and 6 minutes of zoning out with some roughly normally distributed variability with variance around 0.8.*
- This second statement is not strictly speaking true. Like all models, it is a reduction and simplification of reality. However, given the information that we have, it is probably the best description of reality. The hypothesis test was an aid in selecting this model.

Hypotheses vs. Models

- Strictly speaking, the hypothesis test lets us say that: *There is at least one pair of means in our experiment that is not equal.* This is a relatively crude result, but it can be stated with great certainty.
- In contrast, the model we have selected lets us say that: *Given the data collected, we can predict that the mean effects of Apple, Blueberry and Cherry pie dosage on QERM students are about 1, 2 and 6 minutes of zoning out with some roughly normally distributed variability with variance around 0.8.*
- This second statement is not strictly speaking true. Like all models, it is a reduction and simplification of reality. However, given the information that we have, it is probably the best description of reality. The hypothesis test was an aid in selecting this model.

Hypotheses vs. Models: Final Comment

It is often said that all models are wrong, but some can be useful. Perhaps a corrolary might be that hypothesis tests are always accurate (when performed correctly) and always useful, but only for the construction of models - which are all wrong, but occasionally useful.