Analysis of Variance

Eli Gurarie

StatR 101 - Lecture 11a
December 3, 2012

December 3, 2012

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

The Great Pie-eating Zone-out Experiment

An experiment was performed to test the effects of different desserts on student concentration. Twelve (profession and continuing education) 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 all twelve students attended a StatR 101 lecture. 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 ($\bar{x_i}.$)**
Apple Pie	0	2	0.5	1.5		
Blueberry Pie	1	2	3	2		
Cherry Pie	7	5.5	6.5	5		
totals						

The Great Pie-eating Zone-out Experiment

An experiment was performed to test the effects of different desserts on student concentration. Twelve (profession and continuing education) 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 all twelve students attended a StatR 101 lecture. 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:

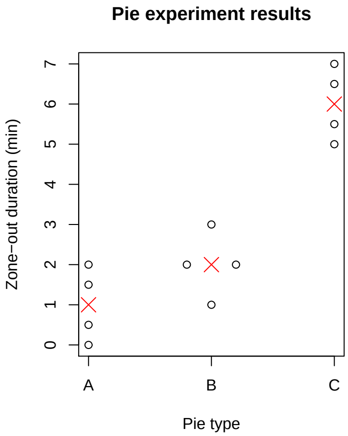
Treatment	**ZOD (min)**				**totals**	**means ($\bar{x_i}.$)**
Apple Pie	0	2	0.5	1.5	4	
Blueberry Pie	1	2	3	2	8	
Cherry Pie	7	5.5	6.5	5	24	
totals					**36**	

The Great Pie-eating Zone-out Experiment

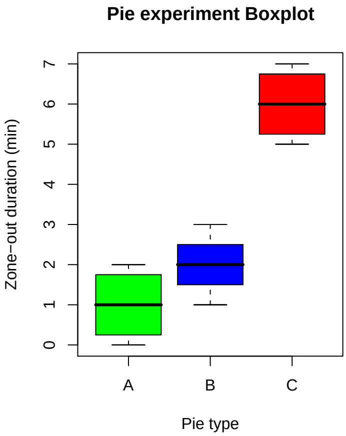
An experiment was performed to test the effects of different desserts on student concentration. Twelve (profession and continuing education) students were divided into three groups of four, each of what was to consume in its entirety an Apple pie, a Blueberry pie, and a Cherry pie. Later all twelve students attended a StatR 101 lecture. 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**

Visualizing the data



Pie experiment results



Pie experiment Boxplot

Formulating a hypothesis

- Research Question:

  Does pie-type affect concentration in students?

- Null hypothesis - in words:

  Different pie types have NO influence on ZOD

- Alternate hypothesis - in words:

  Different pie types DO have influence on ZOD

- Null hypothesis - in math:

  $$\mu_A = \mu_B = \mu_C$$

- Alternate hypothesis - in math:

  $$\mu_A \neq \mu_B \text{ OR } \mu_A \neq \mu_C \text{ OR } \mu_B \neq \mu_C$$

Formulating a hypothesis

- Research Question:

  Does pie-type affect concentration in students?

- Null hypothesis - in words:

  Different pie types have NO influence on ZOD

- Alternate hypothesis - in words:

  Different pie types DO have influence on ZOD

- Null hypothesis - in math:

  $$\mu_A = \mu_B = \mu_C$$

- Alternate hypothesis - in math:

  $$\mu_A \neq \mu_B \text{ OR } \mu_A \neq \mu_C \text{ OR } \mu_B \neq \mu_C$$

Formulating a hypothesis

- Research Question:

  Does pie-type affect concentration in students?

- Null hypothesis - in words:

  Different pie types have NO influence on ZOD

- Alternate hypothesis - in words:

  Different pie types DO have influence on ZOD

- Null hypothesis - in math:

  $$\mu_A = \mu_B = \mu_C$$

- Alternate hypothesis - in math:

  $$\mu_A \neq \mu_B \text{ OR } \mu_A \neq \mu_C \text{ OR } \mu_B \neq \mu_C$$

Comments on models and hypotheses

- So far, we've formulated scientific questions in terms of hypotheses and hypothesis tests (z-tests and t-tests) to compare samples drawn from a population.
- 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 often propose several possible **statistical models** and assess which has greater explanatory power given the quality of the data. The hypothesis test is a *tool* in the *model selection process*.
- This is reflected in the nomenclature. Even a very simple design like the pie experiment, where there is just one more group than the two-sample t-test, we use an *ANALYSIS* of variance, whereas the t-test is 'just' a *TEST*.

Comments on models and hypotheses

- So far, we've formulated scientific questions in terms of hypotheses and hypothesis tests (z-tests and t-tests) to compare samples drawn from a population.
- 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 often propose several possible **statistical models** and assess which has greater explanatory power given the quality of the data. The hypothesis test is a *tool* in the *model selection process*.
- This is reflected in the nomenclature. Even a very simple design like the pie experiment, where there is just one more group than the two-sample t-test, we use an *ANALYSIS* of variance, whereas the t-test is 'just' a *TEST*.

Comments on models and hypotheses

- So far, we've formulated scientific questions in terms of hypotheses and hypothesis tests (z-tests and t-tests) to compare samples drawn from a population.
- 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 often propose several possible **statistical models** and assess which has greater explanatory power given the quality of the data. The hypothesis test is a *tool* in the *model selection process*.
- This is reflected in the nomenclature. Even a very simple design like the pie experiment, where there is just one more group than the two-sample t-test, we use an *ANALYSIS* of variance, whereas the t-test is 'just' a *TEST*.

Formulating a model

- Model 1 - One mean for all groups: $X_{ij} = \mu + \epsilon_{ij}$
- Model 2 - Unique mean in each group: $X_{ij} = \mu_i + \epsilon_{ij}$

Where:

- X_{ij} in one measurement.
- $i \in \{1, 2 ... a\}$ index of the *treatment groups*. Here: $a = 3$, (pies A, B C).
- $j \in \{1, 2 ... n\}$ index of the individual measurement within each group. Here: $n = 4$, $N = a \times n = 12$:
- μ the true total population mean;
- μ_i the true means within each group;
- ϵ_{ij} the random bit of error: **residual**

Important assumption: ϵ_{ij} are **independent**, and are **identically distributed** (iid) with a normal distribution. $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$.

Formulating a model

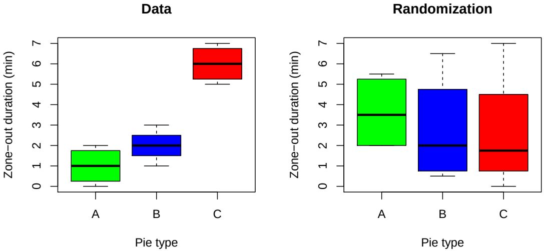
- Model 1 - One mean for all groups: $X_{ij} = \mu + \epsilon_{ij}$
- Model 2 - Unique mean in each group: $X_{ij} = \mu_i + \epsilon_{ij}$

Where:

- X_{ij} in one measurement.
- $i \in \{1, 2...a\}$ index of the *treatment groups*. Here: $a = 3$, (pies A, B C).
- $j \in \{1, 2...n\}$ index of the individual measurement within each group. Here: $n = 4$, $N = a \times n = 12$:
- μ the true total population mean;
- μ_i the true means within each group;
- ϵ_{ij} the random bit of error: **residual**

Important assumption: ϵ_{ij} are **independent**, and are **identically distributed** (iid) with a normal distribution. $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$.

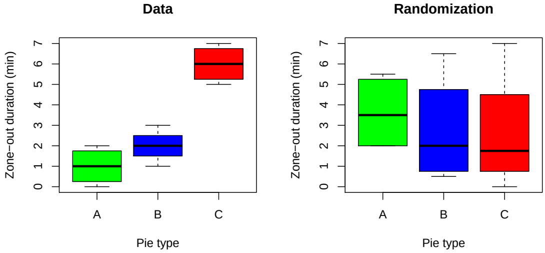
Analysis of variance - ANOVA



The idea behind ANOVA is to compare the variance *within groups* S_i^2 (i.e. the more highly specified model) with *the overall variance* S^2 (i.e. the less specified model).

If S_i^2 is much smaller than S^2, then the treatments have some explanatory power, i.e. a significant amount of total variability is accounted for by the treatment effect.

Analysis of variance - ANOVA



Tr.	means ($\bar{X}_{i\cdot}$)	variances (S_i^2)
A	1	0.8333
B	2	0.6667
C	6	0.8333
total	$\bar{X} = 3.0$	$S^2 = 5.7272$

Tr.	means ($\bar{X}_{i\cdot}$)	variances (S_i^2)
A	3.5	7.1667
B	2.0	6.1667
C	2.5	5.6667
total	$\bar{X} = 3.0$	$S^2 = 5.7272$

Obviously, in our experiment, the variance within groups (left table) is much smaller than the total variance. If we completely randomize our values, the effect vanishes. Our task is to *test* this observation with statistics.

Theory behind ANOVA

1. **Sums of squares:**
   A measure of the total variability in our data is given by the **total sum of squares**:

$$SS_{total} = \sum_{i,j=1}^{N} (X_{ij} - \overline{X})^2$$

2. **Decomposition of the sum of squares:**
3. **Mean sums of squares of treatment and error:**
4. **Distribution of the ratio of MSG and MSE:**
5. **F-test and p-value:**
   Comparing F_0 (test statistic) with $\mathcal{F}_{a-1,N-a}$ gives the p-value of the test. If there is NO treatment effect, we expect F_0 to be around 1. If there is a treatment effect (Null Hypothesis false), then F_0 will be much greater than 1.

Theory behind ANOVA

1. **Sums of squares:**
   A measure of the total variability in our data is given by the **total sum of squares**:

$$SS_{total} = \sum_{i,j=1}^{N} (X_{ij} - \overline{X})^2$$

2. **Decomposition of the sum of squares:**
   The sum of squared can be decomposed into a sum of squares between groups and a sum of squared within groups:

$$SS_{total} = SS_{group} + SS_{errors}$$

$$\begin{aligned} SS_{group} &= n \sum_{i=1}^{a} (\overline{X}_{i\cdot} - \overline{X})^2 : \quad \text{(sum of squares of treatment)} \\ SS_{error} &= \sum_{i=1}^{a} \sum_{j=1}^{n} (X_{ij} - \overline{X}_{i\cdot})^2 : \quad \text{(sum of squares of errors)} \end{aligned}$$

3. **Mean sums of squares of treatment and error:**
4. **Distribution of the ratio of MSG and MSE:**
5. **F-test and p-value:**
   Comparing F_0 (test statistic) with $\mathcal{F}_{a-1,N-a}$ gives the p-value of the test. If

Theory behind ANOVA

1. **Sums of squares:**
2. **Decomposition of the sum of squares:**

$$SS_{group} = n\sum_{i=1}^{a}(\overline{X}_{i\cdot} - \overline{X})^2 : \quad \text{(sum of squares of treatment)}$$

$$SS_{error} = \sum_{i=1}^{a}\sum_{j=1}^{n}(X_{ij} - \overline{X}_{i\cdot})^2 : \quad \text{(sum of squares of errors)}$$

3. **Mean sums of squares of treatment and error:**

$$MS_{group} = SS_{group}/(a-1) : \quad \text{(mean square of group - MSG)}$$

$$MS_{error} = SS_{error}/(N-a) : \quad \text{(mean square of error - MSE)}$$

   BOTH of these are unbiased estimates of σ^2 under the null distribution.

4. **Distribution of the ratio of MSG and MSE:**
5. **F-test and p-value:**
   Comparing F_0 (test statistic) with $\mathcal{F}_{a-1,N-a}$ gives the p-value of the test. If there is NO treatment effect, we expect F_0 to be around 1. If there is a treatment effect (Null Hypothesis false), then F_0 will be much greater than 1.

Theory behind ANOVA

1. **Sums of squares:**
2. **Decomposition of the sum of squares:**
3. **Mean sums of squares of treatment and error:**

$$\begin{aligned} MS_{group} &= SS_{group}/(a-1): \text{ (mean square of group - MSG)} \\ MS_{error} &= SS_{error}/(N-a): \text{ (mean square of error - MSE)} \end{aligned}$$

   BOTH of these are unbiased estimates of σ^2 under the null distribution.

4. **Distribution of the ratio of MSG and MSE:**
   Finally, $F_0 = MS_{group}/MS_{error}$ is a test statistic which *under the null hypothesis* has a known distribution:

$$\frac{MS_{group}}{MS_{error}} \sim \mathcal{F}_{a-1,N-a}$$

5. **F-test and p-value:**
   Comparing F_0 (test statistic) with $\mathcal{F}_{a-1,N-a}$ gives the p-value of the test. If there is NO treatment effect, we expect F_0 to be around 1. If there is a treatment effect (Null Hypothesis false), then F_0 will be much greater than 1.

Theory behind ANOVA

1. **Sums of squares:**
2. **Decomposition of the sum of squares:**
3. **Mean sums of squares of treatment and error:**
4. **Distribution of the ratio of MSG and MSE:**
   Finally, $F_0 = MS_{group}/MS_{error}$ is a test statistic which *under the null hypothesis* has a known distribution:

$$\frac{MS_{group}}{MS_{error}} \sim \mathcal{F}_{a-1,N-a}$$

5. **F-test and p-value:**
   Comparing F_0 (test statistic) with $\mathcal{F}_{a-1,N-a}$ gives the p-value of the test. If there is NO treatment effect, we expect F_0 to be around 1. If there is a treatment effect (Null Hypothesis false), then F_0 will be much greater than 1.

ANOVA table

Thankfully, all of that fits into a single plug-and-chug table:

Source of Variation	**Sum of Squares**	**Degrees of freedom**	**Mean Square**	F_0	p-value
Treatment	SS_{group}	$a - 1$	MSG	$\frac{MSG}{MSE}$	$Pr[\mathcal{F}_{a-1,N-a} > F_0]$
Error	SS_{error}	$N - a$	MSE		
Total	SS_{total}	$N - 1$			

You fill out the table, and at the end all you have to do is look at the *p*-level.

R-code: ANOVA

```
> anova(lm(ZOD ~ Pie))
```

ANOVA table

Thankfully, all of that fits into a single plug-and-chug table:

Source of Variation	**Sum of Squares**	**Degrees of freedom**	**Mean Square**	F_0	p-value
Treatment	SS_{group}	$a-1$	MSG	$\frac{MSG}{MSE}$	$Pr[\mathcal{F}_{a-1,N-a} > F_0]$
Error	SS_{error}	$N-a$	MSE		
Total	SS_{total}	$N-1$			

You fill out the table, and at the end all you have to do is look at the p-level.

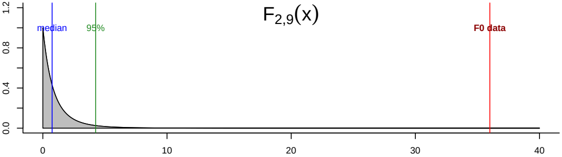
R-code: ANOVA

```
> anova(lm(ZOD ~ Pie))
```

Example of an ANOVA table

ANOVA table of the pie experiment data:

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



The value F_0 is clearly extreme! Calculate the p-value:

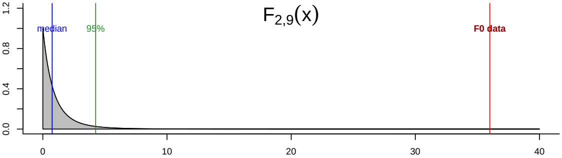
$$\Pr[\mathcal{F}_{2,9} > F_0] = 5.081 \times 10^{-05}$$

The p-value is tiny, so we reject the null hypothesis and conclude that pie-type has a significant effect on zone-out duration.

Example of an ANOVA table

ANOVA table of the pie experiment data:

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



The value F_0 is clearly extreme! Calculate the *p*-value:

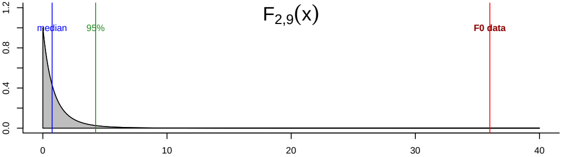
$$\Pr[\mathcal{F}_{2,9} > F_0] = 5.081 \times 10^{-05}$$

The *p*-value is tiny, so we reject the null hypothesis and conclude that pie-type has a significant effect on zone-out duration.

Example of an ANOVA table

ANOVA table of the pie experiment data:

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



The value F_0 is clearly extreme! Calculate the *p*-value:

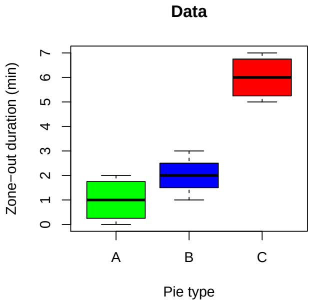
$$\Pr[\mathcal{F}_{2,9} > F_0] = 5.081 \times 10^{-05}$$

The *p*-value is tiny, so we reject the null hypothesis and conclude that pie-type has a significant effect on zone-out duration.

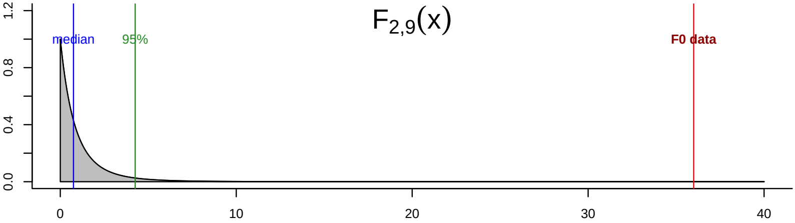
Comparing means of multiple groups.

Question:	Is μ_1 equal to μ_2?	Are μ_1, μ_2, ... μ_n equal?
Test:	Two sample t-test (equal variance)	One way ANOVA
Data:	$\overline{X_1}, \overline{X_2}, n_1 = n_2$, $s_p^2 = \frac{s_1^2+s_2^2}{2}$	$\overline{X_1}, \overline{X_2}, \ldots, \overline{X_a}$, SS_{group}, SS_{error}, $a \times n = N$
Assumptions:	X_1, X_2 ... all normal, iid (equal variance!)	
H_0:	$\mu_1 = \mu_2$	$\mu_1 = \mu_2 = \ldots = \mu_n$
H_A:	$\mu_1 \neq \mu_2$	$\mu_i \neq \mu_j$ for some i and j
Test statistic:	$t_0 = \frac{\overline{X_1}-\overline{X_2}}{\sqrt{2s_p^2/n}}$	$F_0 = \frac{SS_{group}/a-1}{SS_{error}/N-a} = \frac{MSG}{MSE}$
Distribution:	$\mathcal{T}_{2n-2}$	$\mathcal{F}_{a-1,N-a}$
P-value:	$2\,\mathrm{P}(T_{2n-2} > \lvert t_{test} \rvert)$	$\mathrm{P}(\mathcal{F}_{a-1,N-a} > F_0)$

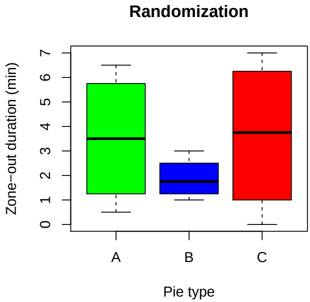
Comparing ANOVA tables



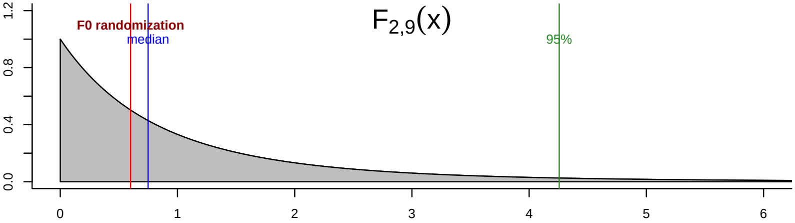
	SS	df	MS	F_0	$\Pr(> F_0)$
Pie	56	2	28	36	5.0e-05
Residuals	7	9	0.78		
Total	63	11			



Comparing ANOVA tables



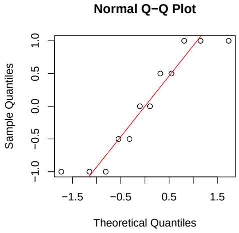
	SS	df	MS	F_0	$\Pr(> F_0)$
Pie	7.6	2	3.8	0.62	0.56
Residuals	55.4	9	6.15		
Total	63	11			

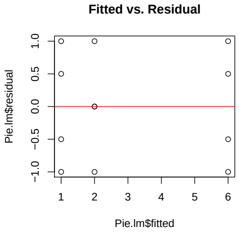


Keep in mind the assumptions of ANOVA:

- The errors are identical in all the groups,
- The errors are independent of each other,
- The errors have a normal distribution

- These assumptions are particularly important because the F-test is not very robust.
- These assumptions are usually tested with the help of **diagnostic plots**.


Normal Q−Q Plot


Fitted vs. Residual

Model specification

Remember our models:

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

ANOVA helped us choose the best model (Model 2). It suggested that if we take into account treatment groups, the σ will be much smaller than if we ignore them.

Now that we have *chosen* a model, we can *specify* it. Our model has 4 parameters: μ_1, μ_2, μ_3 and σ^2. The estimated values for these parameters are:

parameter	estimate	value
μ_1 (Apple pie)	$\overline{X_1}$	1
μ_2 (Blueberry pie)	$\overline{X_2}$	2
μ_3 (Cherry pie)	$\overline{X_3}$	6
σ^2	MS_{error}	0.778

Model specification

Remember our models:

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

ANOVA helped us choose the best model (Model 2). It suggested that if we take into account treatment groups, the σ will be much smaller than if we ignore them.

Now that we have *chosen* a model, we can *specify* it. Our model has 4 parameters: μ_1, μ_2, μ_3 and σ^2. The estimated values for these parameters are:

parameter	estimate	value
μ_1 (Apple pie)	$\overline{X_1}$	1
μ_2 (Blueberry pie)	$\overline{X_2}$	2
μ_3 (Cherry pie)	$\overline{X_3}$	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

Famous posulate:

It is often said that all models are wrong, but some are occasionally useful.

Proposed corollary:

Hypothesis tests are always right (when performed correctly) and always useful, but only for the construction of models - which are always wrong, but occasionally useful.

Hypotheses vs. Models: Final Comment

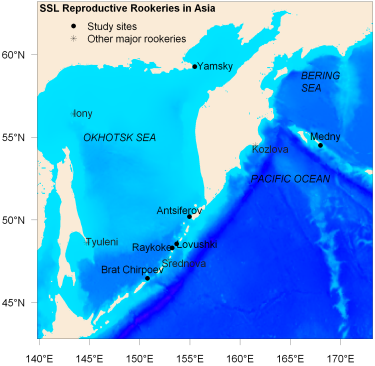
Famous posulate:

It is often said that all models are wrong, but some are occasionally useful.

Proposed corollary:

Hypothesis tests are always right (when performed correctly) and always useful, but only for the construction of models - which are always wrong, but occasionally useful.

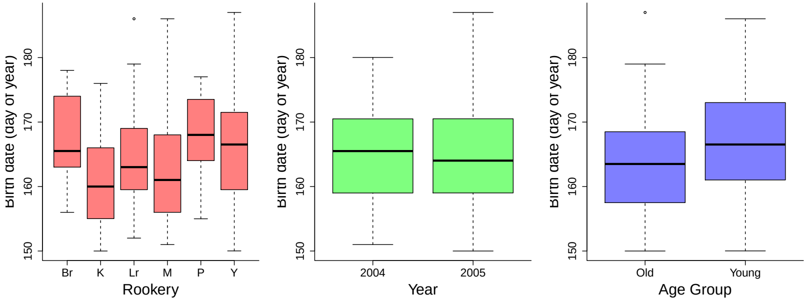
Real example: Birth dates of Steller sea lions



In 2004 and 2005, researchers on 6 reproductive rookeries in the north Pacific observed sea lion pups be born. THe dates females gave birth were observed for 20 felames on each rookery, of which 10 were young (<7 years) and 10 were older (≥ 7 years).

Question: Did the average birth day vary between rookeries, years, and age group of mother?

Box plots



Hard to see any patterns just looking at the boxplots!

Results

ANOVA table:

	d.f.	Sum of Squares	Mean Squares	F-value	Pr(>F)
Island	5	749.74	149.95	2.48	0.0359 *
Age	1	235.76	235.76	3.90	0.0507 *
Year	1	1.88	1.88	0.03	0.8605
Residuals	112	6767.77	60.43		

Results of the analysis indicate that there is a significant difference among **Islands** and, possibly among **Age Groups**, but none between **Years**.

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 (one of) 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".