Comparing means with ANOVA

Material: DeGroot and Schervish 9.7, 11.6
OpenStatistics Chapter 7.5

Slides adopted from Openintro.org

Research question:

You want to test if drinking different beverages affects your reaction time.

You give split your subjects in 3 groups.
You give each group water, tea, and coffee, respectively
You measure their reaction time.

Scenario 1:

		
29	17	10
29	18	11
30	19	12
31	19	12
31	20	13

Scenario 1:

29	17	10
29	18	11
30	19	12
31	19	12
31	20	13

You have little variablity within each group, but different groups look different.

Scenario 2:

		
10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

Scenario 2:

10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

You have lots of variablity within each group, but different groups look the same.

Research question

Is there a difference between the mean response time among the three beverages?

Research question

Is there a difference between the mean response time among the three beverages?

- To compare means of 2 groups we use a Z or a T statistic

Research question

Is there a difference between the mean response time among the three beverages?

- To compare means of 2 groups we use a Z or a T statistic
- To compare means of 3+ groups we use a new test called *ANOVA* and a new statistic called F

ANOVA

Figure out how much of the total variance comes from:

a) The variance between the groups

b) The variance within the groups

Calculate the ratio:

$$F = \frac{\text{variance between groups}}{\text{variance within groups}}$$

The F distributions

Definition: Let Y and W be independent random variables such that

- Y has the χ^2 distribution with m degrees of freedom and
- W has the χ^2 distribution with n degrees of freedom,

where m and n are positive integers.

Then the random variable $X = \frac{Y/m}{W/n}$ follows an F-distribution with m and n degrees of freedom.

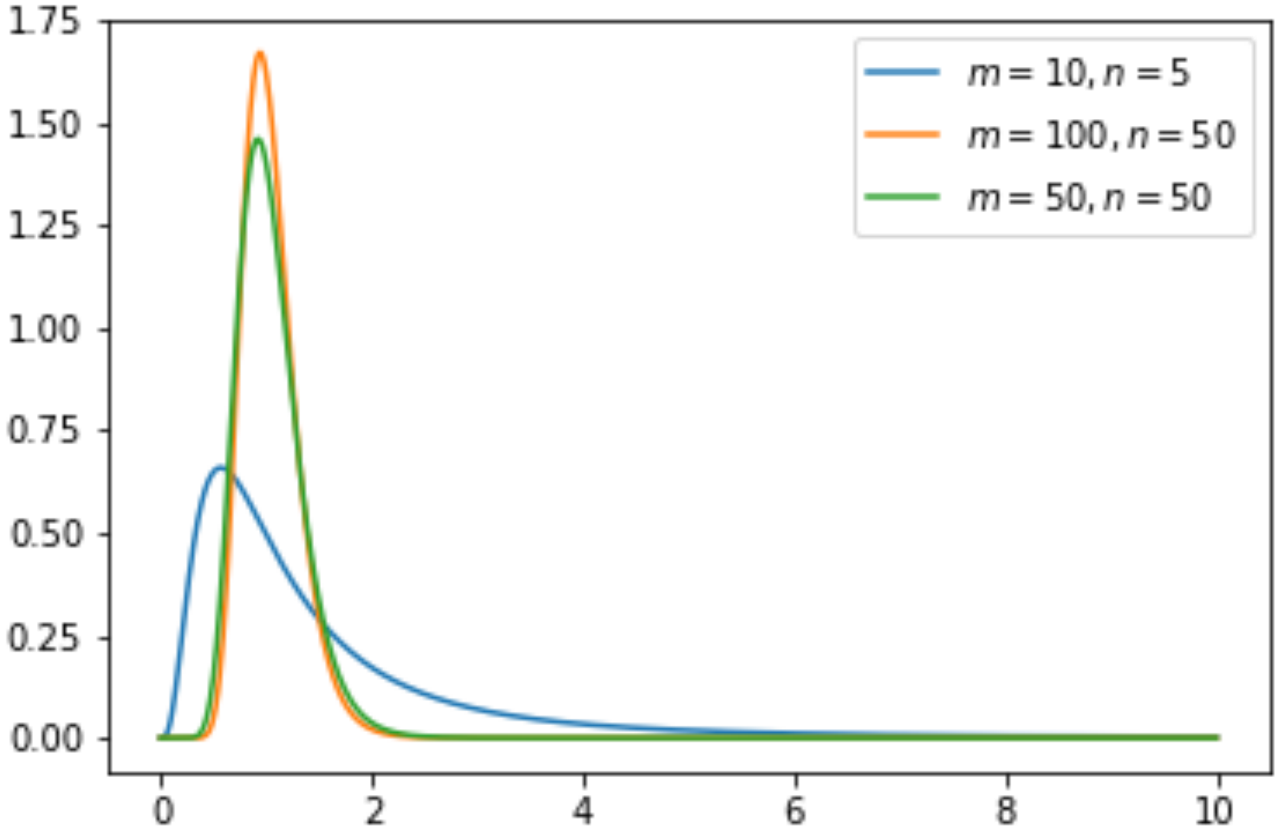
The F distributions

pdf:

$$f(x) = \frac{\Gamma\left[\frac{1}{2}(m+n)\right] m^{\frac{m}{2}} n^{\frac{n}{2}}}{\Gamma\left(\frac{1}{2}m\right)\Gamma\left(\frac{1}{2}n\right)} \times \frac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}, \qquad x > 0$$

Python scipy.stats.f
Quantile using f.ppf

The F distributions



Comparing variances of two normals

Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$

Then $\frac{S_x^2}{\sigma_x^2} \sim \chi^2_{n-1}$, where $S_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$

Comparing variances of two normals

Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$

Then $\frac{S_x^2}{\sigma_x^2} \sim \chi_{n-1}^2$, where $S_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$

Let V $= \frac{S_x^2/[(m-1)\sigma_1^2]}{S_y^2/[(n-1)\sigma_2^2]}$

Then $V \sim F$ distribution with $m-1, n-1$ degrees of freedom.

If $\sigma_1^2 = \sigma_2^2$, then $V = \frac{S_x^2/(m-1)}{S_y^2/(n-1)}$ also follows the same distribution

ANOVA

Figure out how much of the total variance comes from:

a) The variance between the groups

b) The variance within the groups

Calculate the ratio:

$$F = \frac{\text{variance between groups}}{\text{variance within groups}}$$

ANOVA

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable

z/*t* test vs. ANOVA - Purpose

_z_/_t_ test

Compare means from two groups to see whether they are so far apart that the observed difference cannot reasonably be attributed to sampling variability

$$H_0 : \mu_1 = \mu_2$$

ANOVA

Compare the means from two or more groups to see whether they are so far apart that the observed differences cannot all reasonably be attributed to sampling variability

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

ANOVA

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable

H_0: The mean outcome is the same across all categories,

$$\mu_1 = \mu_2 = \ldots = \mu_k,$$

where μ_i represents the mean of the outcome for observations in category i

ANOVA

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable

H_0 : The mean outcome is the same across all categories,

$$\mu_1 = \mu_2 = \ldots = \mu_k,$$

where μ_i represents the mean of the outcome for observations in category *i*

H_A : At least one mean is different than others

Hypotheses

A. $H_0 : \mu_W = \mu_T = \mu_C$
$H_A : \mu_W \neq \mu_T \neq \mu_C$

B. $H_0 : \mu_W \neq \mu_T \neq \mu_C$
$H_A : \mu_W = \mu_T = \mu_C$

C. $H_0 : \mu_W = \mu_T = \mu_C$
H_A : At least one mean is different

A. $H_0 : \mu_W = \mu_T = \mu_C = 0$
H_A : At least one mean is different

E. $H_0 : \mu_W = \mu_T = \mu_C$
$H_A : \mu_B > \mu_M > \mu_C$

Hypotheses

A. $H_0 : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}}$
$H_A : \mu_{\mathrm{W}} \neq \mu_{\mathrm{T}} \neq \mu_{\mathrm{C}}$

B. $H_0 : \mu_{\mathrm{W}} \neq \mu_{\mathrm{T}} \neq \mu_{\mathrm{C}}$
$H_A : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}}$

C. $H_0 : \mu_W = \mu_T = \mu_C$
H_A : *At least one mean is different*

A. $H_0 : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}} = 0$
H_A : At least one mean is different

E. $H_0 : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}}$
$H_A : \mu_{\mathrm{B}} > \mu_{\mathrm{M}} > \mu_{\mathrm{C}}$

Data

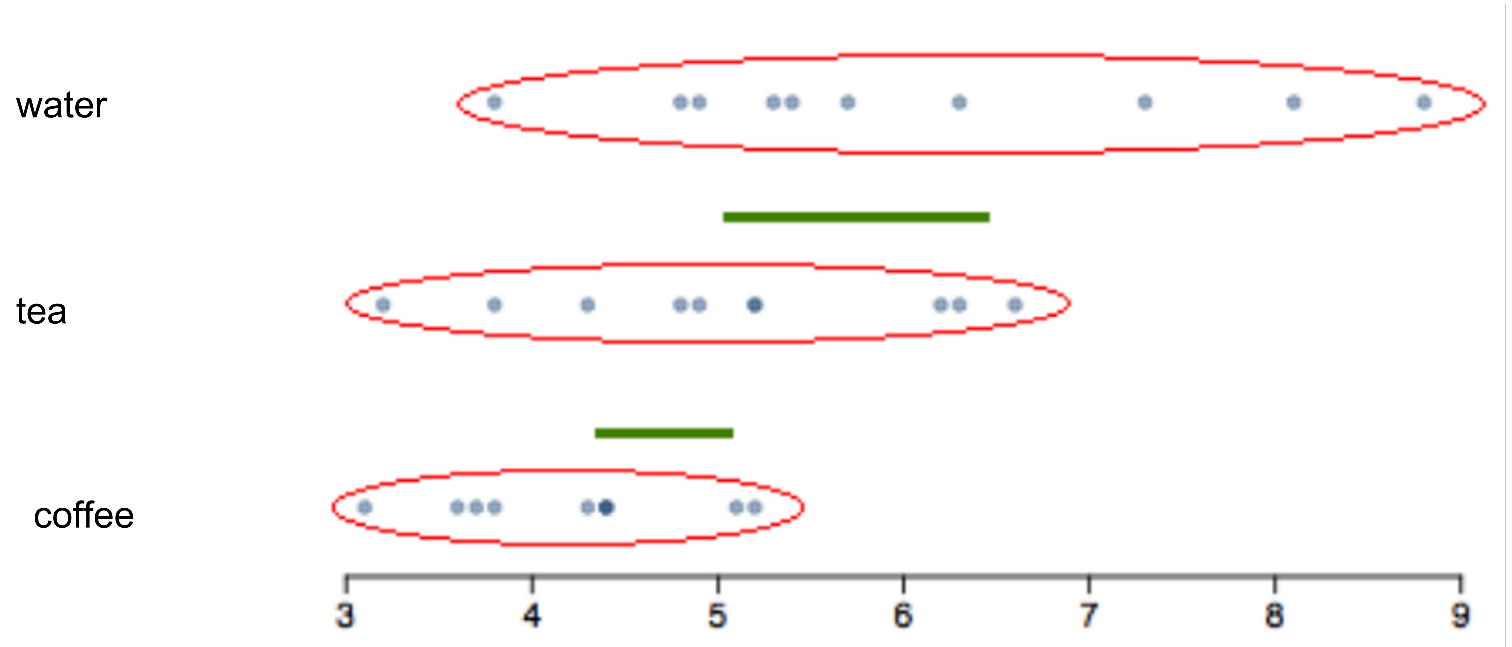
Reaction times

	Time(sec)	beverage
1	3.80	water
2	4.80	water
...		
10	8.80	water
11	3.20	tea
12	3.80	tea
...		
20	6.60	tea
21	3.10	coffee
22	3.60	coffee
...		
30	5.20	coffee

Test statistic

Does there appear to be a lot of variability within groups? How about between groups?

$$F = \frac{variability\ bet.\ groups}{variability\ within\ groups}$$



Measuring variability

Total:

$$SST = \sum_{i} (x_i - \bar{x})^2$$

Between Groups:

$$SSG = \sum_{i=1}^{k} n_i (\bar{x_i} - \bar{x})^2$$

Residual:

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x_i}\right)^2$$

Measuring variability

Total:

$$SST = \sum_{i} (x_i - \bar{x})^2$$

Between Groups:

$$SSG = \sum_{i=1}^{p} n_i (\overline{x_i} - \bar{x})^2$$

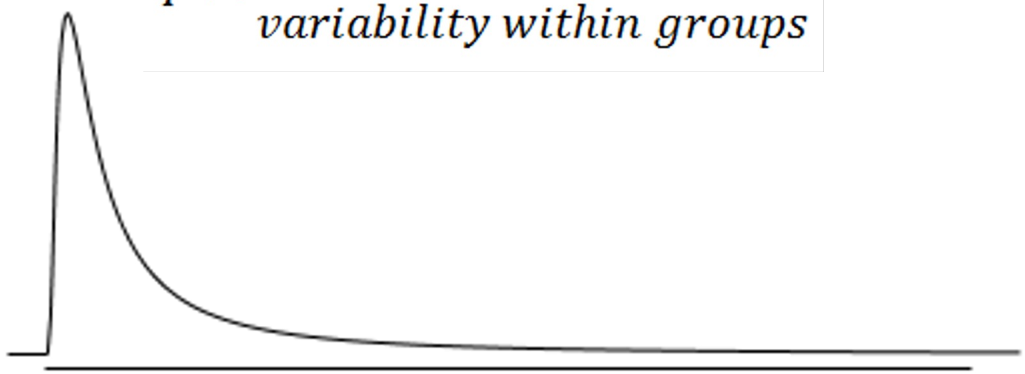
Residual:

$$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x_i}\right)^2$$

$$SST = SSG + SSE$$

F distribution and p-value

$$F = \frac{\text{variability bet. groups}}{\text{variability within groups}}$$



- Large values of the F statistic lead to small p-values, which leads to rejecting In order to be able to reject H_0, we need a small p-value, which requires a large F statistic
- In order to obtain a large F statistic, variability between sample means needs to be greater than variability within sample means

Theorem

Suppose $\mu_1 = \mu_2 = \cdots = \mu_k$ and

$$\sigma_1 = \sigma_2 = \cdots = \sigma_k$$

Then

$$F = \frac{SSG/(k-1)}{SSE/(n-k)}$$

has the F distribution with $k-1$ and $n-k$ degrees of freedom

(Data i.i.d, $X^j \sim N\left(\mu_j, \sigma_j^2\right) for\ j = 1, \ldots, k$)

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean

	n	mean
water	10	6.04
tea	10	5.05
coffee	10	4.2
overall	30	5.1

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean

	n	mean
water	10	6.04
tea	10	5.05
coffee	10	4.2
overall	30	5.1

$$SSG = (10 \times (6.04 - 5.1)^2)$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean

	n	mean
water	10	6.04
tea	10	5.05
coffee	10	4.2
overall	30	5.1

$$SSG = (10 \times (6.04 - 5.1)^2) + (10 \times (5.05 - 5.1)^2)$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean

	n	mean
water	10	6.04
tea	10	5.05
coffee	10	4.2
overall	30	5.1

$$\begin{aligned} SSG = {} & (10 \times (6.04 - 5.1)^2) \\ & + (10 \times (5.05 - 5.1)^2) \\ & + (10 \times (4.2 - 5.1)^2) \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean

	n	mean
water	10	6.04
tea	10	5.05
coffee	10	4.2
overall	30	5.1

$$SSG = (10 \times (6.04 - 5.1)^2) + (10 \times (5.05 - 5.1)^2) + (10 \times (4.2 - 5.1)^2)$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean

	n	mean
water	10	6.04
tea	10	5.05
coffee	10	4.2
overall	30	5.1

$$\begin{aligned} SSG &= (10 \times (6.04 - 5.1)^2) \\ &+ (10 \times (5.05 - 5.1)^2) \\ &+ (10 \times (4.2 - 5.1)^2) \\ &= 16.96 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares total, SST

Measures the variability between groups

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where x_i represent each observation in the dataset

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares total, SST

Measures the variability between groups

$$SST = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

where x_i represent each observation in the dataset

$SST = (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \ldots + (5.2 - 5.1)^2$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares total, SST

Measures the variability between groups

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where x_i represent each observation in the dataset

$$SST = (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \ldots + (5.2 - 5.1)^2$$
$$= (-1.3)^2 + (-0.3)^2 + (-0.2)^2 + \ldots + (0.1)^2$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares total, SST

Measures the variability between groups

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where x_i represent each observation in the dataset

$$\begin{aligned} SST &= (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \ldots + (5.2 - 5.1)^2 \\ &= (-1.3)^2 + (-0.3)^2 + (-0.2)^2 + \ldots + (0.1)^2 \\ &= 1.69 + 0.09 + 0.04 + \ldots + 0.01 \\ &= 54.29 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares error, SSE

Measures the variability within groups:

$$SSE = SST - SSG$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares error, SSE

Measures the variability within groups:

$$SSE = SST - SSG$$

$$SSE = 54.29 - 16.96 = 37.33$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Mean squared error

Mean squared error is calculated as sum of squares divided by the degrees of freedom

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Mean squared error

Mean squared error is calculated as sum of squares divided by the degrees of freedom

$$MSG = 16.96/2 = 8.48$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Mean square error

Mean square error is calculated as sum of squares divided by the degrees of freedom

$$MSG = 16.96/2 = 8.48$$
$$MSE = 37.33/27 = 1.38$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Test statistic, *F* value

As we discussed before, the *F* statistic is the ratio of the between group and within group variability

$$F = \frac{MSG}{MSE}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	*16.96*	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Test statistic, *F* value

As we discussed before, the *F* statistic is the ratio of the between group and within group variability

$$F = \frac{MSG}{MSE}$$

$$F = \frac{8.48}{1.38} = 6.14$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

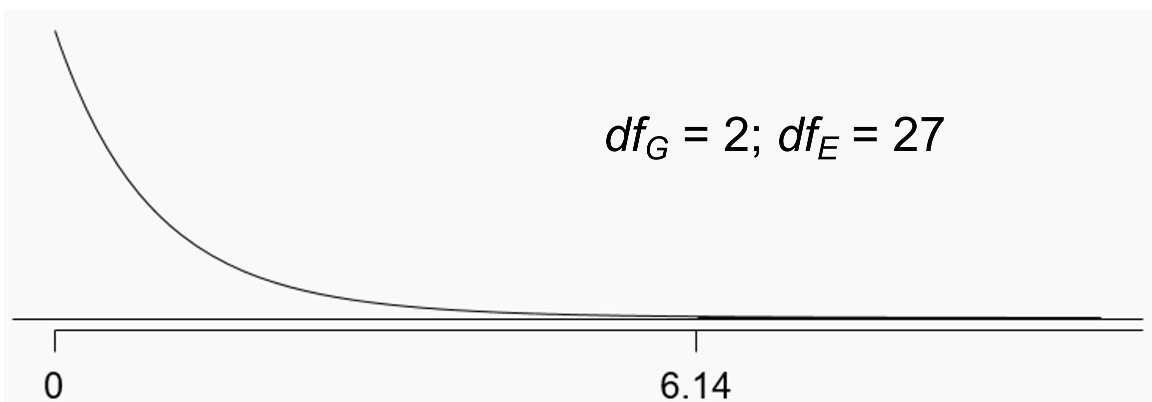
p-value

p-value is the probability of at least as large a ratio between the "between group" and "within group" variability, if in fact the means of all groups are equal. It's calculated as the area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	beverage	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

p-value

p-value is the probability of at least as large a ratio between the "between group" and "within group" variability, if in fact the means of all groups are equal. It's calculated as the area under the *F* curve, with degrees of freedom df_G and df_E, above the observed *F* statistic.



Conclusion - in context

What is the conclusion of the hypothesis test?

The data provide convincing evidence that the average reaction time

A. is different for all beverages
B. with coffee is lower than the other beverages
C. is different for at least one beverage
D. is the same for all beverages

Conclusion - in context

What is the conclusion of the hypothesis test?

The data provide convincing evidence that the average reaction time

A. is different for all beverages
B. with coffee is lower than the other beverages
C. is different for at least one beverage
D. is the same for all beverages

Conclusion

- If p-value is small (less than α), reject H_0. The data provide convincing evidence that at least one mean is different from (but we can't tell which one)

Conclusion

- If p-value is small (less than α), reject H_0. The data provide convincing evidence that at least one mean is different from (but we can't tell which one)
- If p-value is large, fail to reject H_0. The data do not provide convincing evidence that at least one pair of means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance)

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied
   - Carefully consider whether the data may be independent (e.g. no pairing)
   - Always important, but sometimes difficult to check

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied
   - Carefully consider whether the data may be independent (e.g. no pairing)
   - Always important, but sometimes difficult to check

2. The observations within each group should be nearly normal
   - Especially important when the sample sizes are small

How do we check for normality?

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied
   - Carefully consider whether the data may be independent (e.g. no pairing)
   - Always important, but sometimes difficult to check

2. The observations within each group should be nearly normal
   - Especially important when the sample sizes are small

How do we check for normality?

3. The variability across the groups should be about equal
   - Especially important when the sample sizes differ between groups

How can we check this condition?

(1)independence

Does this condition appear to be satisfied?

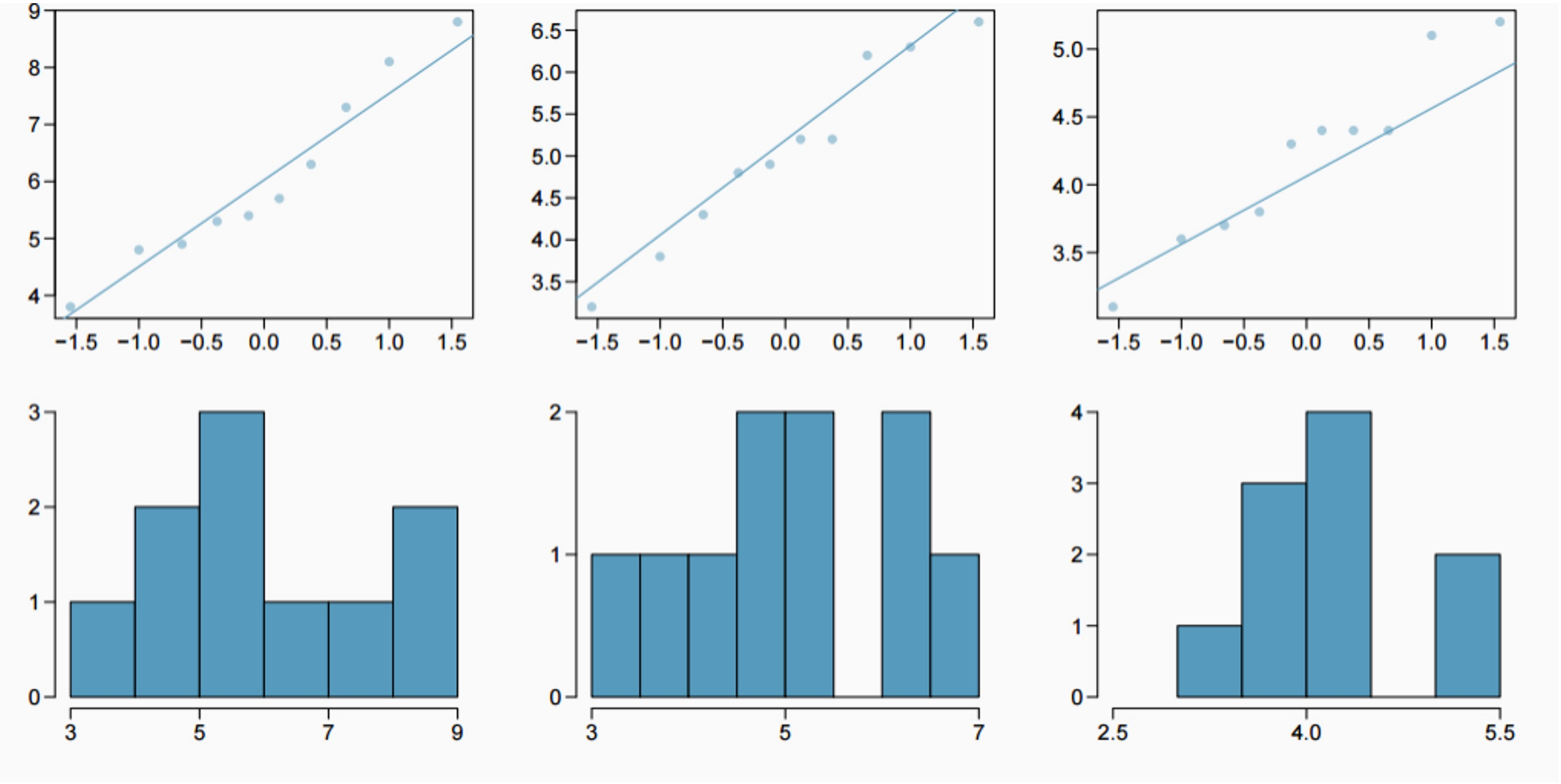
(1)independence

Does this condition appear to be satisfied?

In this study the we have no reason to believe that the condition is not satisfied (i.e., people are randomly chosen)

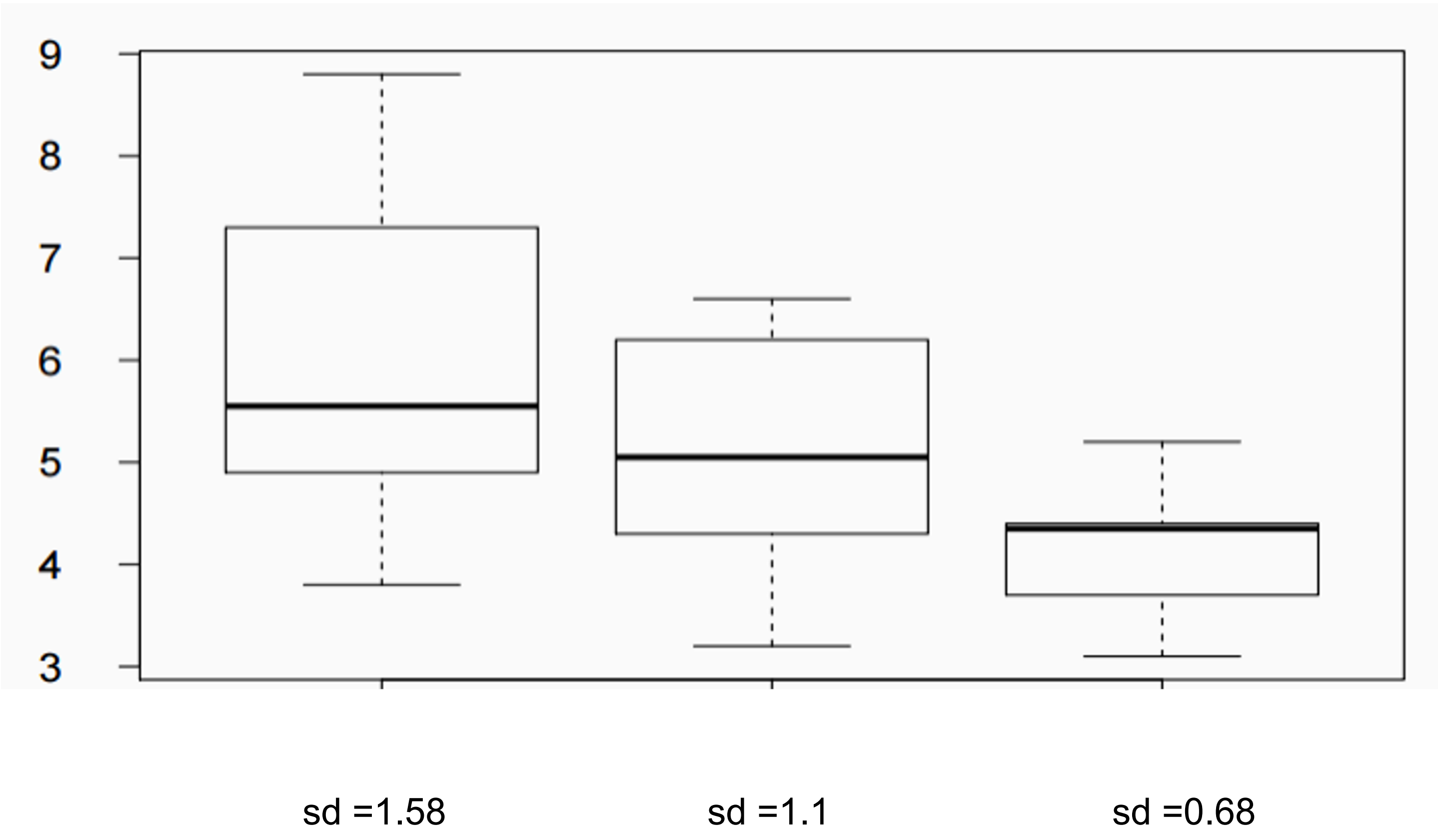
(2)approximately normal

Does this condition appear to be satisfied?



(3)constant variance

Does this condition appear to be satisfied?



Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”
- We can do two sample t tests for differences in each possible pair of groups

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”
- We can do two sample t tests for differences in each possible pair of groups

Can you see any pitfalls with this approach?

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”
- We can do two sample *t* tests for differences in each possible pair of groups

Can you see any pitfalls with this approach?

- When we run too many tests, the Family-wise error rate increases
- We can use: Corrections for multiple comparisons (e.g., Bonferroni)

Why not just use pairwise comparisons?

- Controlling for family-wise error rate is conservative

- It may be the case that we end up getting no significant p-values in pairwise comparisons, but a significant ANOVA p-value

Exercise

Plan 1	Plan 2	Plan 3
3	3.5	8
4.5	7	4
4	4.5	3
3		

Three different diet plans are to be tested for mean weight loss. The entries in the table below are the weight losses for the different plans.

a) Test the hypothesis that all three diet plans have the same mean weight loss by performing ANOVA.
b) Can you think of a way for testing if the assumption of equal variances holds for plan 1 and plan 2?

Exercise

$$SST = \sum_i (x_i - \bar{x})^2$$

Between Groups:

$$SSG = \sum_{i=1}^{k} n_i(\overline{x_i} - \bar{x})^2$$

Residual:

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \overline{x_i})^2$$

Plan 1	Plan 2	Plan 3
3	3.5	8
4.5	7	4
4	4.5	3
3		

$$F = \frac{MSG}{MSE} = \frac{\frac{SSG}{k-1}}{\frac{SSE}{n-k}}$$

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
(Group)	Diet plans					
(Error)	Residuals					
	Total					

Exercise

$$SST = \sum_i (x_i - \bar{x})^2$$

Between Groups:

$$SSG = \sum_{i=1}^{p} n_i(\bar{x_i} - \bar{x})^2$$

Residual:

$$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x_i}\right)^2$$

Plan 1	Plan 2	Plan 3
3	3.5	8
4.5	7	4
4	4.5	3
3		

$$F = \frac{MSG}{MSE} = \frac{\frac{SSG}{k-1}}{\frac{SSE}{n-k}}$$

		Df	Sum sq	Mean sq	F value	$\Pr(> \mathrm{F})$
(Group)	Diet plans	2	4.5375	2.2687	0.7158	0.5214
(Error)	Residuals	7	22.1875	3.1696		
	Total	9	26.725			