IMS22: Analysis of Variance (ANOVA)

Packages

```
library(openintro)
library(tidyverse)
library(infer)
library(broom)
```

Analysis of Variance (ANOVA)

The analysis of variance method compares means of several groups.

- Let k denote the number of groups.
- Each group has a corresponding population of subjects.
- The means of the outcome variable for the k populations are denoted by $\mu_1, \mu_2, \ldots, \mu_k$.

Hypotheses and Assumptions for the ANOVA Test Comparing Means

The analysis of variance is a significance test of the null hypothesis of equal population means:

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

The alternative hypothesis is:

$$H_A : \text{At least one mean is different from another.}$$

Hypotheses and Assumptions for the ANOVA Test Comparing Means

The assumptions for the ANOVA test comparing population means are as follows:

- **Independence**: In a survey sample, independent random samples are selected from each of the k populations. For an experiment, subjects are randomly assigned separately to the k groups.
- **Normality**: The data for each of the k groups are approximately normal. Looking for symmetry, no big outliers.
- **Constant Variance**: The variance in the groups needs to be about equal from one group to the next. Rule of thumb: the largest sample standard deviation should not be more than double the smallest one.

Example

We would like to discern whether there are real differences between the on-base percentage (OBP) of baseball players according to their position: outfielder (OF), infielder (IF), and catcher (C). We will use a dataset called `mlb_players_18`, which includes batting records of 429 Major League Baseball (MLB) players from the 2018 season who had at least 100 at bats. The on-base percentage roughly represents the fraction of the time a player successfully gets on base or hits a home run.

Example

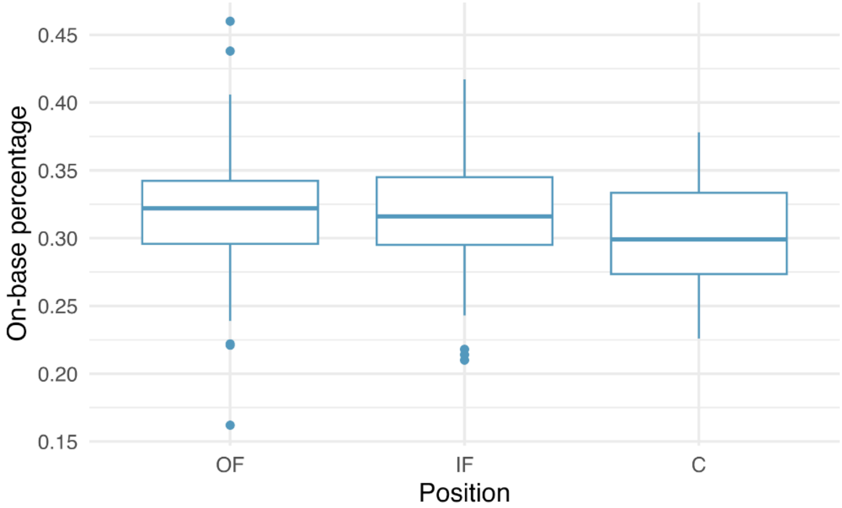
VARIABLE	DESCRIPTION
name	Player name
team	The abbreviated name of the player's team
position	The playerś primary field position (OF, IF, C)
AB	Number of opportunities at bat
H	Number of hits
HR	Number of home runs
RBI	Number of runs batted in
AVG	Batting average, which is equal to H/AB
OBP	On-base percentage, which is roughly equal to the fraction of times a player gets on base or hits a home run

Example

name	team	position	AB	H	HR	RBI	AVG	OBP
Abreu, J	CWS	IF	499	132	22	78	0.265	0.325
Acuna Jr., R	ATL	OF	433	127	26	64	0.293	0.366
Adames, W	TB	IF	288	80	10	34	0.278	0.348
Adams, M	STL	IF	306	73	21	57	0.239	0.309
Adduci, J	DET	IF	176	47	3	21	0.267	0.290
Adrianza, E	MIN	IF	335	84	6	39	0.251	0.301

`position` is the **grouping variable** and `OBP` is the **outcome variable**

Example



0.45
0.40
0.35
0.30
0.25
0.20
0.15
On-base percentage
OF
IF
C
Position

Short cut?

The largest difference between the sample means is between the catcher and the outfielder positions. Consider again the original hypotheses:

$$H_0 : \mu_{IF} = \mu_{OF} = \mu_C$$

H_A : At least one mean is different from another.

Can we run the test by simply estimating whether the difference of μ_C and μ_{OF} is 0?

Data Snooping

Can we run the test by simply estimating whether the difference of μ_C and μ_{OF} is 0?

This is called **data snooping** or **data fishing**. This would leading to an inflation in the Type 1 Error rate, and a invalid procedure. The primary issue here is that we are inspecting the data before picking the groups that will be compared.

This is related to the **prosecutor's fallacy**.

ANOVA

The ANOVA method is used to compare population means from many groups *simultaneously*.

It is called analysis of variance because it uses evidence about two types of variability.:

- **mean square between groups (MSG)**: a measure of the variability between the groups; with associated degrees of freedom $df_G = k - 1$
- **mean square error (MSE)**: a measure of the variability within the groups; with associated degrees of freedom $df_E = n - k$

ANOVA F Test Statistic

The analysis of variance (ANOVA) F test statistic summarizes:

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

The larger the variability *between* groups relative to the variability *within* groups, the larger the F test statistic tends to be.

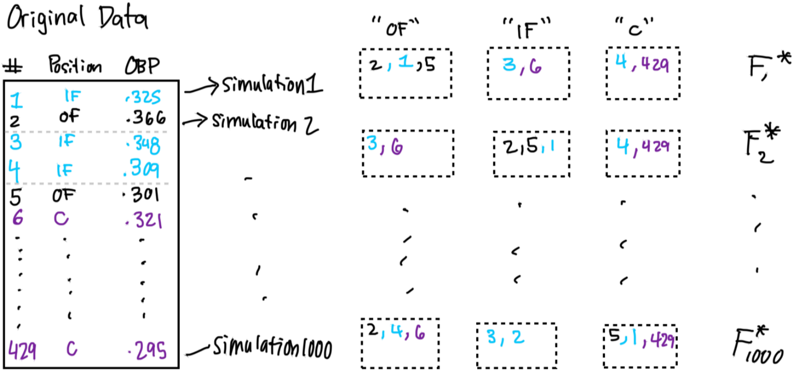
ANOVA F Test Statistic

The test statistic for comparing means has the F sampling distribution.

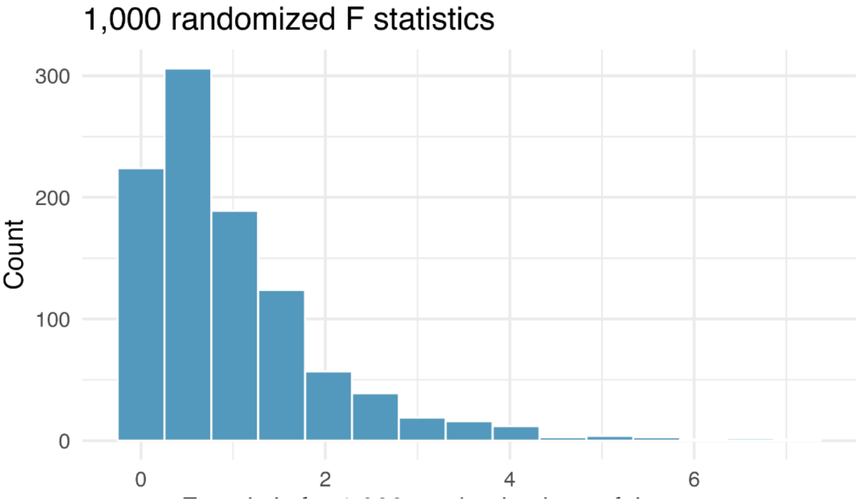
- randomization test
- mathematical model

The larger the F test statistic value, the stronger the evidence against H_0.

Randomization Test with the F-Distribution

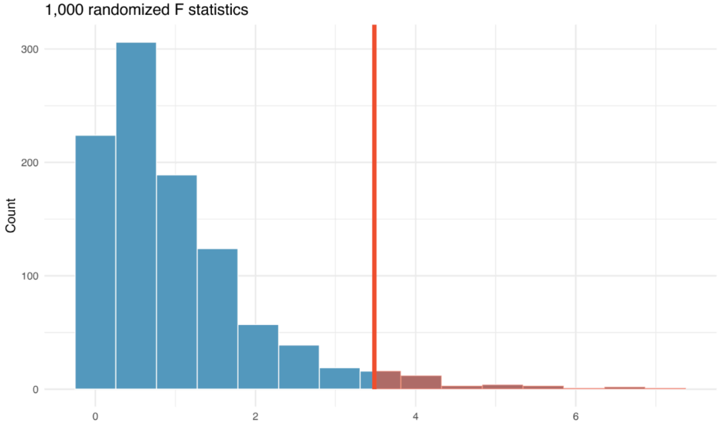


Randomization Test with the F-Distribution



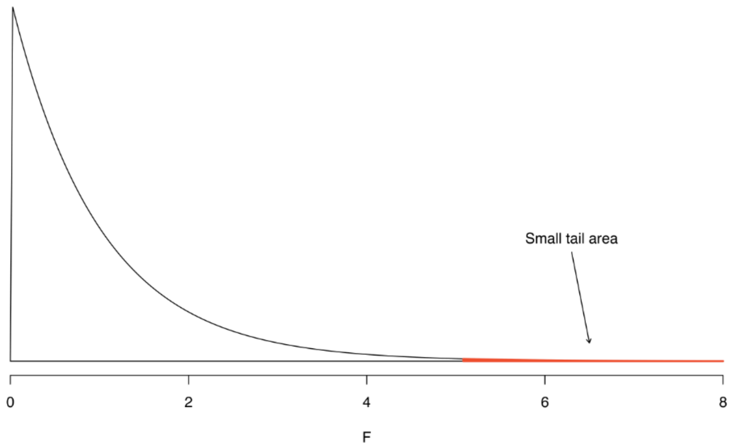
*Caution: slightly different data.

Randomization Test with the F-Distribution



P-value is *always* a right-tail area!

Mathematical Model for F-Distribution

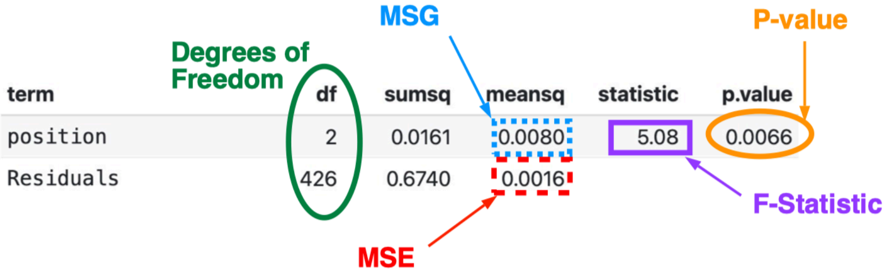


Same idea, but use mathematical properties (instead of the computer) to see what the possible values for F would be if H_0 was true.

Mathematical Model for F-Distribution

We can get the p-value and other relevant statistics from an ANOVA table in R.

term	df	sumsq	meansq	statistic	p.value
position	2	0.0161	0.0080	5.08	0.0066
Residuals	426	0.6740	0.0016		



term	df	sumsq	meansq	statistic	p.value
position	2	0.0161	0.0080	5.08	0.0066
Residuals	426	0.6740	0.0016		

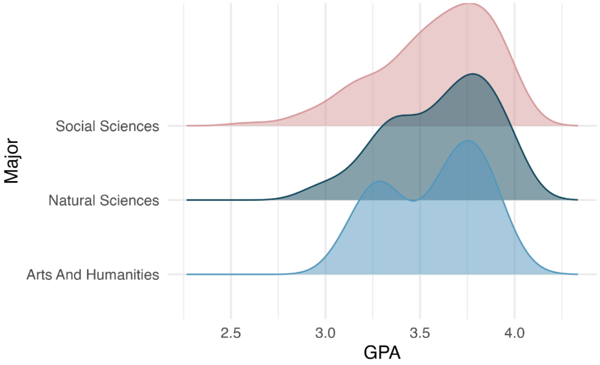
SUMMARY: ANOVA F test for Comparing Population Means of Several Groups

1. Frame the research question in terms of hypotheses.
2. Collect data (check conditions)
   - Independence, Normality for each group, same variance for each group.
3. Model the randomness that would occur if the null hypothesis was true.
   - Randomization Test, or use the F distribution
4. Analyze the data.
   - Calculate the P-Value (no confidence interval option). Always a right tail probability.
5. Form a conclusion.
   - The F-test does not tell us which groups differ or how different they are.
   - All we know is at least one group mean is different from the rest.

Practice Question

[IMS 22.9] GPA and major.. Undergraduate students taking an introductory statistics course at Duke University conducted a survey about GPA and major. The plots show the distribution of GPA among three groups of majors. Also provided is the ANOVA output.

a) Write the hypotheses for testing for a difference between average GPA across majors.
b) Do you think the conditions are met? Explain.



Practice Question

[IMS 22.9] GPA and major.

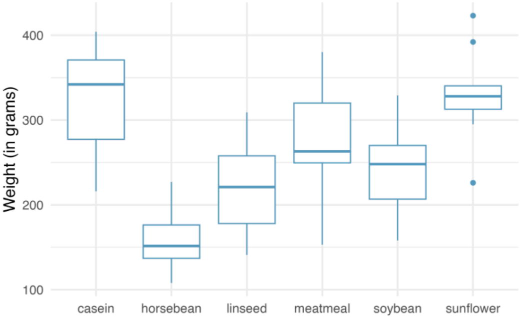
term	df	sumsq	meansq	statistic	p.value
`major`	2	0.03	0.02	0.21	0.81
`Residuals`	195	15.77	0.08		

c) What is the conclusion of the hypothesis test?

d) How many students answered these questions on the survey, i.e. what is the sample size?

Practice Question

[IMS 22.5] Chicken diet and weight, many groups. An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Sample statistics and a visualization of the observed data are shown below.



Feed type	Mean	SD	n
casein	323.58	64.43	12
horsebean	160.20	38.63	10
linseed	218.75	52.24	12
meatmeal	276.91	64.90	11
soybean	246.43	54.13	14
sunflower	328.92	48.84	12

Practice Question

[IMS 22.5] Chicken diet and weight, many groups.

Preview first 12 rows in data set

```
head(chickwts, n = 12)
```

```
   weight      feed
1     179 horsebean
2     160 horsebean
3     136 horsebean
4     227 horsebean
5     217 horsebean
6     168 horsebean
7     108 horsebean
8     124 horsebean
9     143 horsebean
10    140 horsebean
11    309   linseed
12    229   linseed
```

Practice Question

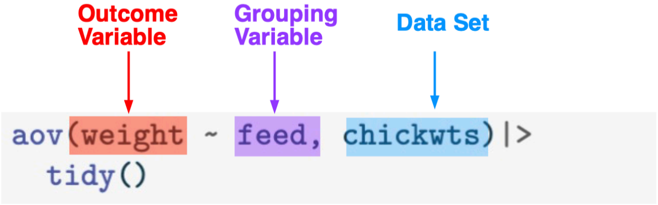
[IMS 22.5] Chicken diet and weight, many groups.

ANOVA Table in R

```
aov(weight ~ feed, chickwts)|>
  tidy()
```

```
# A tibble: 2 x 6
  term         df   sumsq meansq statistic   p.value
  <chr>     <dbl>   <dbl>  <dbl>     <dbl>     <dbl>
1 feed          5 231129. 46226.      15.4  5.94e-10
2 Residuals    65 195556.  3009.      NA    NA
```

[IMS 22.5] Chicken diet and weight, many groups.



Practice Question

[IMS 22.5] Chicken diet and weight, many groups.

```
# A tibble: 2 x 6
  term         df   sumsq meansq statistic   p.value
  <chr>     <dbl>   <dbl>  <dbl>     <dbl>     <dbl>
1 feed          5 231129. 46226.      15.4  5.94e-10
2 Residuals    65 195556.  3009.      NA    NA
```

Conduct a hypothesis test to determine if these data provide convincing evidence that the average weight of chicks varies across some (or all) groups. Make sure to check relevant conditions.

Practice Question

[IMS 22.7] Coffee, depression, and physical activity. Participants in a study investigating the relationship between coffee consumption and exercise were asked to report the number of hours they spent per week on exercise. Based on these data the researchers estimated the total hours of metabolic equivalent tasks (MET) per week, a value always greater than 0. The table below gives summary statistics of MET for women in this study based on the amount of coffee consumed.

	Caffeinated coffee consumption				
	1 cup / week or fewer	**2-6 cups / week**	**1 cups / day**	**2-3 cups / day**	**4 cups / day or more**
Mean	18.7	19.6	19.3	18.9	17.5
SD	21.1	25.5	22.5	22.0	22.0
n	12,215.0	6,617.0	17,234.0	12,290.0	2,383.0

a) Write the hypotheses for evaluating if the average physical activity level varies among the different levels of coffee consumption.

b) Check conditions and describe any assumptions you must make to proceed with the test.

Practice Question

[IMS 22.7] Coffee, depression, and physical activity.

c) Below is the output associated with this test. What is the conclusion of the test?

	df	sumsq	meansq	statistic	p.value
cofee	4	10,508	2,627	5.2	0
Residuals	50,734	25,564,819	504		
Total	50,738	25,575,327			