Analysis of Variance

553.211 Intro to Prob. & Stats.

Johns Hopkins University

EN.553.211

What is Analysis of Variance?

- Let's step back for a moment to our two independent samples t-test (or z-test for large samples). The purpose of this test was to decide whether or not two population means were equal

$$H_0 : \mu_1 = \mu_2 \quad \text{against} \quad H_a : \mu_1 \neq \mu_2.$$

Question

What if we are sampling from more than two populations?

What NOT to do

Let's suppose for a moment that we are sampleing from 3 normal populations with the same variance $\mathcal{N}(\mu_1, \sigma)$, $\mathcal{N}(\mu_2, \sigma)$, $\mathcal{N}(\mu_3, \sigma)$.
We have the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ against the alternative that $\mu_i \neq \mu_j$ for at least one pair $i \neq j$.
We first propose the strategy:

1. Conduct a two sample t-test of $H_{0,a} : \mu_1 - \mu_2 = 0$ against $H_{a,1} : \mu_1 - \mu_2 \neq 0$ at sig. lv. α.
2. Conduct a two sample t-test of $H_{0,b} : \mu_1 - \mu_3 = 0$ against $H_a : \mu_1 - \mu_3 \neq 0$ at sig. lv. α.
3. Conduct a two sample t-test of $H_{0,c} : \mu_2 - \mu_3 = 0$ against $H_a : \mu_2 - \mu_3 \neq 0$ at sig. lv. α.

Rule: Reject H_0 if we reject at least one of $H_{0,a}, H_{0,b}, H_{0,c}$.
Reasonable right? What is the problem?

What NOT to do Continued

The probability of a type I error with this strategy is

$$\begin{aligned} P(\text{reject } H_0 \mid H_0 \text{ true}) &= P(\text{reject at least one } H_{0,a}, H_{0,b}, H_{0,c} \mid \text{true}) \\ &= 1 - P(\text{reject none of } H_{0,a}, H_{0,b}, H_{0,c} \mid \text{true}) \\ &= 1 - (1-\alpha)^3 \end{aligned}$$

In general, if we have k populations, this strategy gives a significance level of $1 - (1-\alpha)^{\binom{k}{2}}$, which is very bad!
We need a better technique for testing whether or not the means of several populations are equal, which is called **Analysis of variance (ANOVA)**.

When is ANOVA used?

Let's look as an experiment which would be a typical situation in which ANOVA would be appropriate:

- A group of people participating in an experiment is randomly divided into:
  - Those who did not eat breakfast.
  - Those who had a partial breakfast.
  - Those who had a full breakfast.

  Every participant takes an aptitude test and their scores are recorded. We are interested in whether or not eating breakfast affects the score.

When is ANOVA Used? Continued

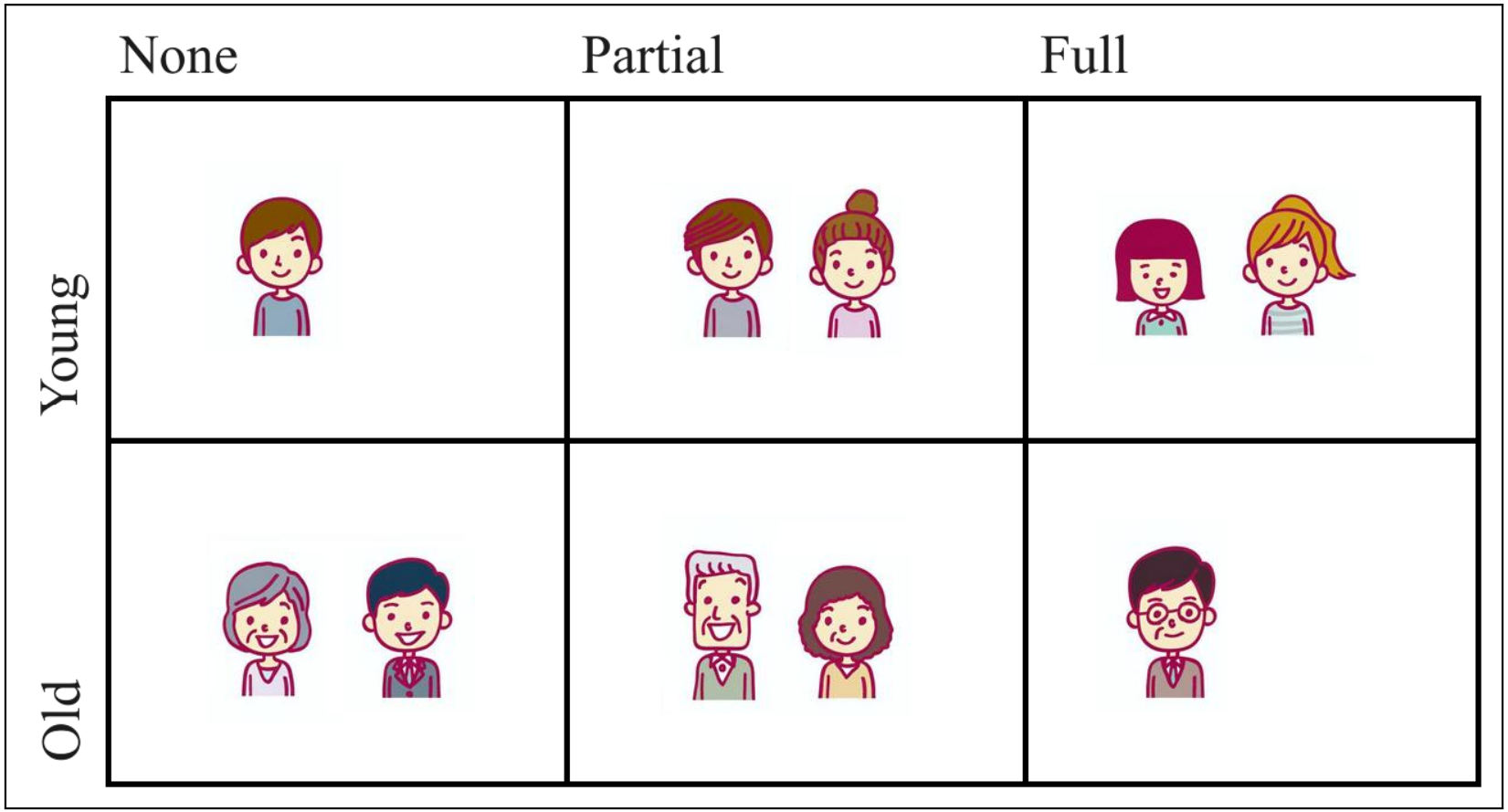
In this example we would say

- The participants of the study are the **experimental units.**
- Eating breakfast is a **factor** of the experiment.
- The amount of breakfast (none, partial, full) are the **levels** of the breakfast factor.
- Their test score is the **response** being measured.

Question

What if we were to split the participants again by age? So we might have young people with no breakfast, old people with no breakfast, young people with partial breakfast, etc..

When is ANOVA Used? Continued



When is ANOVA Used? Continued

In this case, we now have two factors:

- The breakfast factor (with three levels).
- The age factor (with two levels).
- Any specific combination of factor levels is called a **treatment**. E.g. young full breakfast.

For right now, will only focus experiments with a *single* factor, called **one-way ANOVA**. We will see later how to generalize our analysis.

The Right Hypothesis

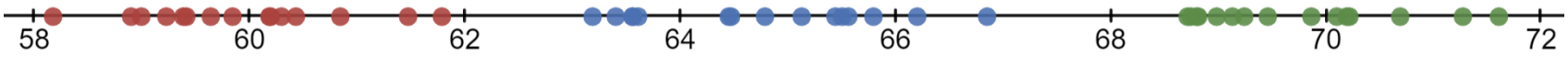
If we are sampling from k independent populations, we need a way to test the hypothesis

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ against $H_a : \mu_i \neq \mu_j$ (least one pair $i \neq j$).

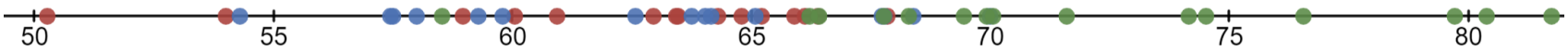
ANOVA will give us a way to do this.

An Intuitive Idea:

Scenario A- We take a sample of 15 from each of three normally distributed populations (red, blue, and green) with equal variances:



Scenario B- We take a sample of 15 from each of three populations, also normally distributed with equal variances colored red,blue, and green:



Which scenario gives more evidence of at least one population having different mean?

The F Statistic

To compare multiple population means simultaneously, we need a test statistic which will measure

1. **variation between the sample means**. E.g. the sample means of the red, blue and green samples are far apart in scenario A, and
2. **natural variation within the samples**. E.g. the red, blue and green observations within each sample are all close together with their respective color.

Although there is a lot of computational work that goes into the formula, ultimately, the test statistic we use in ANOVA is the ratio of the above two things:

$$F = \frac{\text{variation between the sample means}}{\text{variation within the samples}}$$

The larger the observed value of F is, the more evidence we have that at least one of the means is different from the others.

Computing the F Statistic

First suppose we take a random sample from each of k normal populations with the same variance.

Let $x_{i,j}$ be the j^{th} measurement in the i^{th} sample. That is

Pop. 1 Data	**Pop. 2 Data**	...	**Pop. k Data**
$x_{1,1}$	$x_{2,1}$	$\cdots$	$x_{k,1}$
$x_{1,2}$	$x_{2,2}$	$\cdots$	$x_{k,2}$
$\vdots$	$\vdots$		$\vdots$
x_{1,n_1}	x_{2,n_2}	$\cdots$	x_{k,n_k}

Here, each n_i is the size of the sample from the i^{th} population. We'll denote the *overall* sample size by $N = n_1 + n_2 + \cdots + n_k$.

Computing the F Statistic; Continued

The details leading to the F statistic are presented in six steps, ending with an ANOVA table.

- Calculate the mean and variance for each sample: $\overline{x}_i$ and s_i^2.
- Calculate the overall sample mean

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{k} n_i \overline{x}_i$$

- Calculate the **sum of squares between treatments**[1]

$$SST := \sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x})^2.$$

- Calculate the **sum of squares within treatments due to error**

$$SSE := \sum_{i=1}^{k} (n_i - 1) s_i^2.$$

[1] You may also see this called sum of squares for groups, SSG

Computing the F Statistic; Continued

- Calculate **total sum of squares**

$$\text{Total } SS := SST + SSE.$$

Finally we fill in the ANOVA table:

Source	df	SS	MS	F
Treatments	$k-1$	SST	$MST = \frac{SST}{(k-1)}$	MST/MSE
Error	$N-k$	SSE	$MSE = \frac{SSE}{(N-k)}$	
Total	$N-1$	Total SS		

ANOVA Table; Source Column, Treatments Row

Remember, the F statistic is the ratio of variation between sample means and the variation within each sample. Thus we have two **sources** of variation.

The first source of variation is summarized in the **Treatments** row. This is the variation among the sample means. Which explains why we used the formula

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

- n_i = sample size from population i.
- $\overline{x}_i$ = sample mean from population i.
- $\overline{x}$ = overall sample mean.

ANOVA Table; Source Column, Treatments Row Cont.

Fact

If $H_0 : \mu_1 = \cdots = \mu_k$ were true, then SST/σ^2 follows the χ^2_{k-1} distribution.

Intuitively, H_0 being true means that the overall sample mean $\bar{X}$ is an unbiased estimator of *every* treatment sample mean $\bar{X}_i$. So that SST is basically[2] a sample variance of k normal RVs $\bar{X}_1, \ldots, \bar{X}_k$; which we know from Chapter 10 follows the χ^2 distributed with $df = (\text{sample size}) - 1$.

[2] It is actually more complicated than this, but not necessary to understand the precise details in this class.

ANOVA Table; Source Column, Error Row

The second source of variation is summarized in the **Error** row. This characterizes the natural variation within the samples themselves. Which is why we use the formula

$$SSE = \sum_{i=1}^{k} (n_i - 1) s_i^2$$

- $n_i - 1 =$ degrees of freedom in sample i.
- $s_i^2 =$ sample variance from population i.

Fact

SSE/σ^2 follows the χ^2_{N-k} distribution. Note this is always, true regardless of H_0.

ANOVA Table; *df* Column

Because the SST measures the variation among the k sample means, we have that the SST has $k-1$ degrees of freedom.

On the other hand, we have $n_i - 1$ degrees of freedom for each of the i samples we took. Thus the df in the error column should be

$$\sum_{i=1}^{k}(n_i - 1) = \underbrace{\sum_{i=1}^{k} n_i}_{\text{total } \#} - \underbrace{\sum_{i=1}^{k} 1}_{\# \text{ of samples}}$$
$$= N - k.$$

ANOVA Table; *SS* Column

The **sum of squares** (SS) column records the sum of squares between treatments, SST, and the sum of squares due to error, SSE. The sum of this column gives the total sum of squares Total SS.

Fact

$$\text{Total } SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2.$$

In words, the sum $SST + SSE$ is the sum of squares over all observations in all the samples.

So you could calculate SSE by first finding the Total SS, then subtracting SST. (Sometimes this is easier, depending on the information given to you.)

ANOVA Table; *MS* Column

The **mean squares** column is the ratio of the entries in the *SS* column to the entries in the *df* column.

So the **mean squares between treatments** is

$$MST = \frac{SST}{k-1}$$

while the **mean squares due to error** is

$$MSE = \frac{SSE}{N-k}.$$

ANOVA Table; The F Statistic

Finally, the ANOVA table leads to the F statistic

$$F = \frac{MST}{MSE}.$$

Fact

If $H_0 : \mu_1 = \cdots = \mu_k$ were true, then

$$F = \frac{MST}{MSE} = \frac{SST/(k-1)}{SSE/(N-k)} \sim F_{k-1,N-k}$$

since it is the quotient of two independent χ^2 RVs divided by their respective degrees of freedom. Notice it was necessary to assume equal variances so that the σ^2's cancel in the numerator and denominator.

Comment about the null-hypothesis

Remark: Some texts present the RVs in each sample i as

$$X = \mu + \alpha_i + \varepsilon,$$

where μ is the pooled population mean (a parameter), α_i is the *effect* that treatment i has on the pooled mean μ (also a parameter), while ε is a mean zero normally distributed RV with the same variance for all $i = 1, 2, \ldots, k$. Then our hypothesis of is equivalent to that of

$$H_0 : \mu_1 = \cdots = \mu_k \iff H_0 : \alpha_1 = \cdots = \alpha_k = 0.$$

It is occasionally easier to think of an ANOVA in terms of the latter; as the variation in X explained by the treatment is given by α_i and is estimated by MST, while the variation explained by error is given by $V(\varepsilon) = \sigma^2$ and is estimated by MSE.

Example 1

Let's go back to **Scenario A** and **Scenario B** from the beginning of lecture.

Remember we thought that **Scenario A** gave more evidence for different means, since there was no "mixing".

Let's compare the ANOVA table from each scenario, and see if the F statistics agree with our initial guess.

Example 1; Continued

Scenario A				
Source	*df*	*SS*	*MS*	*F*
Treatments	2	720.82	360.41	345.97
Error	42	43.75	1.0417	
Total	44	764.57		

Scenario B				
Source	*df*	*SS*	*MS*	*F*
Treatments	2	849.18	424.59	14.97
Error	42	1190.92	28.35523	
Total	44	2040.10		

Discussion of Assumptions

- Just like with the student's t-distribution and the F-distribution we learned before, we need to assume we are random sampling from normal populations with a common variance σ^2.
- Just like previous weeks, the analysis of variance procedures we will learn this week are fairly **robust**.
- That is, as long as sample sizes are roughly the same size, and the population histograms are roughly "bell-shaped", the test-statistics' distribution will not differ significantly from theory.

Example 2

Example (Fill in the Blanks)

Complete the ANOVA table with the missing values

Source	*df*	*SS*	*MS*	*F*
Treatments	?	?	?	?
Error	?	47.4	2.9625	
Total	19	123.2		

Solution

We can find the df of the SSE by

$$MSE = \frac{SSE}{df} \implies 2.9625 = \frac{47.4}{df}.$$

Thus, $df = 16$.

This immediately gives us $df = 3$ for the SST. Of course we can find SST by

$$\text{Total } SS = SST + SSE \implies 123.2 = 47.4 + SST.$$

Thus, $SST = 75.8$.This implies

$$MST = \frac{SST}{df} = \frac{75.8}{3} \approx 25.267 \quad \text{and} \quad F = \frac{MST}{MSE} = \frac{25.267}{2.9625} \approx 8.53$$

Example 3

Example (One-Way ANOVA Hypothesis Test)

A study is designed to test whether there is a difference in mean daily calcium intake in adults with normal bone density, adults with osteopenia (a low bone density which may lead to osteoporosis) and adults with osteoporosis. The data are shown below.

Normal Bone Density	Osteopenia	Osteoporosis
1200	1000	890
1000	1100	650
980	700	1100
900	800	900
750	500	400
800	700	350

Is there a statistically significant difference in mean calcium intake in patients between categories?

Solution

Let μ_1, μ_2, and μ_3 be the average calcium intake among patients with normal bone density, osteopenia, and osteoporosis, respectively. We will test the hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

against

$$H_a : \mu_i \neq \mu_j \text{ for at least one pair } i \neq j.$$

We were not given a significance level, so we may as well assume $\alpha = 0.05$. Now we need to collect the statistics to create the ANOVA table.

Solution; Continued

Step 1: First compute the sample means for each group:

$$\begin{array}{lll} \overline{x}_1 = 938.3 & \overline{x}_2 = 800.0 & \overline{x}_3 = 715.0 \\ s_1^2 = 26016.67 & s_2^2 = 48000 & s_3^2 = 89950 \end{array}$$

Step 2: Calculate the *overall* sample mean:

$$\begin{aligned} \overline{x} &= \frac{1}{N} \sum_{i=1}^{3} n_i \overline{x}_i \\ &= \frac{1}{18} \left(6 \cdot 1938.3 + 6 \cdot 800.0 + 6 \cdot 715.0 \right) \\ &\approx 817.78 \end{aligned}$$

Solution; Continued

Step 3: Calculate the SST :

$$\begin{aligned} SST &= \sum_{i=1}^{3} n_i(\overline{x}_i - \overline{x})^2 \\ &= 6(938.3 - 817.78)^2 + 6(800.0 - 817.78)^2 + 6(715.0 - 817.78)^2 \\ &\approx 152477.7 \end{aligned}$$

Step 4: Now the SSE :

$$\begin{aligned} SSE &= \sum_{i=1}^{3}(n_i - 1)s_i^2 \\ &= 5(26016.7 + 48000 + 89950) \\ &= 819833.3 \end{aligned}$$

Solution; Continued

Step 5: We can now complete the ANOVA table:

Source	*df*	*SS*	*MS*	*F*
Treatments	2	152477.7	$\frac{152477.7}{2} = 76238.6$	$\frac{76238.6}{54655.5} = 1.395$
Error	15	819833.3	$\frac{819833.3}{15} = 54655.5$	
Total	17	972311.0		

Step 6: Find the critical value[3] of $F_{0.05} \approx 3.68$. Since $1.395 < 3.68$, we retain H_0.

[3] We have numerator $df = 2$ and denominator $df = 15$.

Remarks

Takeaways: Hypothesis testing with ANOVA works in the same way it did before. However, since the nature of the experiment is more complicated, the procedure for computing the test statistic is much more involved as well.

Luckily, there are many statistical programs (Excel, r, MINITAB, SQL,...) which can instantly run ANOVA tests! So in real life, we almost never do this process by hand.

Estimating Difference in Treatments

- Suppose you've run the ANOVA test and found evidence to reject H_0. Then you might have the follow-up question: Which means are different from others?
- You can construct confidence intervals using the small-sample procedures be learned before, based on the Student's t-distribution.

(1-α)100% CI

Fact

Let S^2 be the mean squares due to error in an ANOVA experiment. Then for all $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, k$ (except i)

$$\frac{\bar{X}_i - \mu_i}{S/\sqrt{n_i}} \sim \mathcal{T}_{N-k} \quad \text{and} \quad \frac{(\bar{X}_i - \bar{X}_j) - (\mu_i - \mu_j)}{\sqrt{S^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim \mathcal{T}_{N-k}$$

Question: Why use *MSE* instead of S_i^2? They are both unbiased estimators of σ^2, after all?
Answer: Because *MSE* uses the information from *all* the samples, not just i and/or j. So it gives you a better estimate of σ^2 (again it is important that all populations have equal variance).

(1-α)100% CI Continued

Definition

The $(1-\alpha)100\%$ **confidence interval for a single treatment mean** is

$$\overline{x}_i \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n_i}}$$

The $(1-\alpha)100\%$ **confidence interval for the difference between two treatment means** is

$$(\overline{x}_i - \overline{x}_j) \pm t_{\alpha/2}\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}.$$

In each of the above, $\overline{x}_i$ is the sample mean from population i, while

$$s^2 = MSE = \frac{SSE}{N-k},$$

so $t_{\alpha/2}$ is based on $N-k$ degrees of freedom.

Two-Way Classification (Blocking)

- In many instances it is clear that experimental units are *not* homogeneous. Thus some (or a lot) of the variation in the SSE may be explained by these non-homogeneous subgroups within the treatments.
- There are two ways to study this effect. The first we will look at is the randomized block design.
- The **randomized block design** is a direct extension of the paired-difference design from weeks ago.
- The design partitions the k treatments into b **blocks** that are relatively similar, and randomly assigns one unit within each block to each treatment.

ANOVA Table for Block Design

Source	df	SS	MS	F
Treatments	$k-1$	SST	$MST = \frac{SST}{k-1}$	$\frac{MST}{MSE}$
Blocks	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$	$\frac{MSB}{MSE}$
Error	$(b-1)(k-1)$	SSE	$MSE = \frac{SSE}{(b-1)(k-1)}$	
Total	$bk-1$	Total SS		

What's New?

The biggest addition to this table is the Blocks row. It contains the **degrees of freedom** of the blocks, as well as **sum of squares between blocks**:

$$SSB = \sum_{i=1}^{b} k(\overline{B}_i - \overline{x})^2$$

where $\overline{B}_i$ is the mean within each block $i = 1, 2, \ldots, b$. Again, k is the number of treatments, and $\overline{x}$ is the overall average.

The sum of squares between treatments can now be calculated as

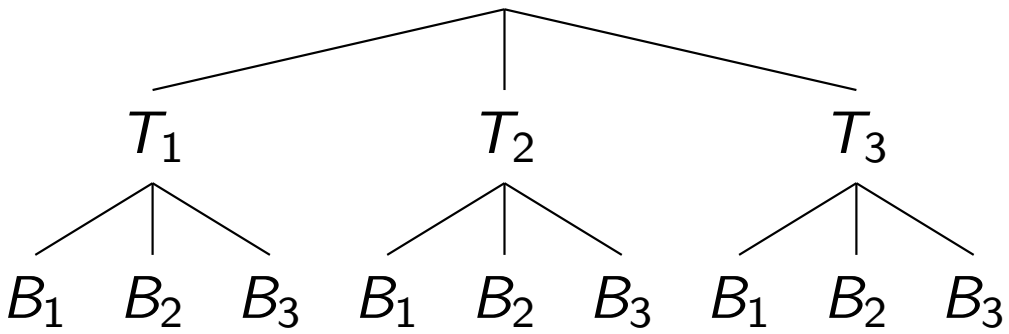
$$SST = \sum_{i=1}^{k} b(\overline{T}_i - \overline{x})^2$$

where $\overline{T}_i$ is the mean within each treatment $i = 1, 2, \ldots, k$.

What's New? Continued

Notice that we now have an additional F statistic MSB/MSE. Unsurprisingly, this is because we can hypothesis test whether there is a difference in means between the treatments *as well as* whether there is a difference in means between the blocks!

In the *df* column of the Error row you'll find $(b-1)(k-1)$, why is this?



There are $k-1$ degrees of freedom in the treatment levels, for each of these there are $b-1$ in the block levels.

What's New? Continued

The sum of squares due to error can be calculated by either

- directly with the formula

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{b} (x_{ij} - \overline{T}_i - \overline{B}_j + \overline{x})^2$$

- or by subtracting the SST from the Total SS

$$SSE = (\text{Total } SS) - SST - SSB.$$

Remember Total SS is the overall sum of squares (you can use the Excel formula $(bk - 1)*$VAR.S to compute this).

Example 4

Example (Cell Phone Costs)

A consumer watchdog group decides to compare the average costs for four cell phone companies (A,B,C, and D) using three different usage levels (Low, Medium, and High) as blocks.

Bl. \ Tr.	**A**	**B**	**C**	**D**	Mean
Low	27	24	31	23	26.25
Medium	68	76	65	67	69
High	308	326	312	300	311.5
Mean	134.33	142	136	130	135.58

Create an ANOVA table for the data.

Solution

We calculate the necessary information:

$$\begin{aligned} SST &= 3 \times \big((134.33 - 135.58)^2 + (142 - 135.58)^2 \\ &\quad + (136 - 135.5)^2 + (130 - 135.58)^2\big) \\ &\approx 222.25 \end{aligned}$$

$$\begin{aligned} SSB &= 4 \times \big((26.25 - 135.58)^2 + (69 - 135.58)^2 \\ &\quad + (311.5 - 135.58)^2\big) \\ &\approx 189335.2 \end{aligned}$$

Finally,

$$\text{Total } SS = 189799.9$$

Solution; Continued

We can now solve

$$SSE \approx 189798.9 - 222.25 - 189335.2 = 241.45$$

Then find the MST, MSB, and MSE by dividing by the degress of freedom:

$$MST = \frac{SST}{k-1} = \frac{222.25}{3} \approx 74.083$$

$$MSB = \frac{189335.2}{2} = 94667.5$$

$$MSE = \frac{241.45}{6} \approx 40.24$$

Solution; Continued

Finally the two F statistics are

- Treatments:

$$F = \frac{MST}{MSE} = \frac{74.083}{40.25} \approx 1.841$$

- Blocks:

$$F = \frac{MSB}{MSE} = \frac{94667.5}{40.25} \approx 2352$$

We can now complete the ANOVA table:

Solution; Continued

Source	*df*	*SS*	*MS*	*F*
Treatments	3	222.25	74.083	1.841
Blocks	2	189335.2	94667.5	2352
Error	6	241.45	40.25	
Total	11	189799.9		

Hypothesis Testing

Notice in example 4 that the F statistic for the treatments (1.841) was pretty small. This corresponds to a P-value[4] of about 0.240.

This should make sense, as the averages between the four companies were not very different. On the other hand, the F statistic for the blocks (2352) is so large, the corresponding P-value is basically 0. This should make sense as well, as more cell phone usage would naturally lead to more charges from the company.

It seems it's the case that almost all the variability seen in cost is explained by the usage level, rather than the different companies. If each company offers a similar product, then this conclusion should agree with our intuition: *If one company really did have a lower average for the same product, then the others could not stay competitive!*

[4] use $df_1 = 3$ in the numerator and $df_2 = 6$ in the denominator.

Cautionary Comments

- A randomized block design should not be used when treatments and blocks both correspond to **experimental** factors of interest.
- In Example 4, we did not have reason to believe that being a "Low Usage" customer of Company A had significant difference than being a Low Usage customer of Companies B, C, or D.
- If we suspect the effect in the response variable will be similar regardless of the block we are using, then a block-design experiment would be appropriate.

Cautionary Comments; Continued

- In the Breakfast, Age, and Test Score example we looked at at the very beginning, we might have reason to think that not having breakfast has a larger impact on test score for young people than for old. If so, then there would be an **interaction effect** between the breakfast and age factor.
- When an interaction is suspected between two factors, you should analyze the data as a **factorial experiment**, which we will introduce next week/lecture.