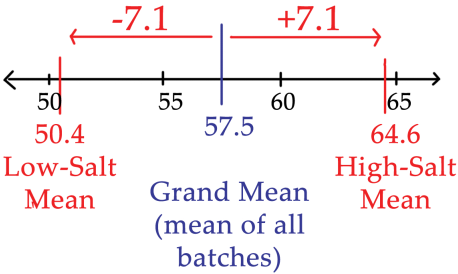
Section 3: ANOVA

William Christensen and H. Dennis Tolley

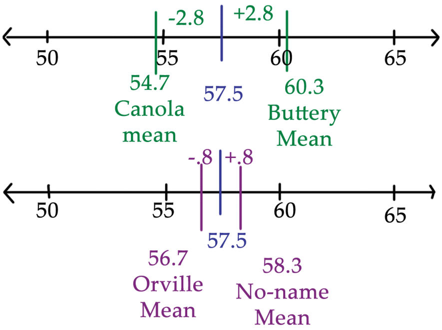
Introduction

Consider a popcorn experiment comparing two levels of salt (low and high), two levels of oil (canola and buttery), and two levels of brand (no-name and Orville).

- Start by comparing mean of the 4 low-salt batches with mean of the 4 high-salt batches (where batch scores were obtained from the average score among all tasters):



Also,



But, means like 50.4 and 64.6 are partly due to the <u>actual</u> (hard-to-measure) tastiness score and partly due to the errors associated with measurements.

- Unfortunately, we don't see what part of a measurement is error... but we can estimate it.

Decomposing a Measurement

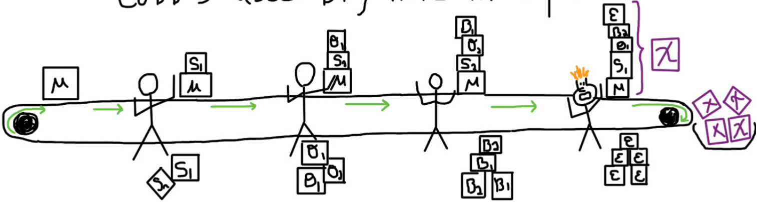
We want to think about measured values having components.
Observed "tastiness" score:

$$\overbrace{x_{ijk}}^{\text{Observed}} = \overbrace{\mu + S_i + O_j + B_k}^{\text{True Value}} + \overbrace{\varepsilon_{ijk}}^{\text{Residual Error}}$$

S_1: low salt O_1: canola oil B_1: No-name brand

S_2: high salt O_2: buttery oil B_2: Orville brand

- Cobb's assembly line metaphor



Suppose a batch of popcorn with

- high salt (S_2)
- Canola oil $(\mathcal{O}_1)$
- No-name brand kernels (B_1) and
- the measurement score is $x_{211} = 64$

If we have the estimates of $\mu, S_2, \mathcal{O}_1$, and B_1, we can estimate the error value ε_{211} :

$$
\begin{aligned}
64 &= \mu + S_2 + \mathcal{O}_1 + B_1 + \hat{\varepsilon}_{211} \\
&= 57.5 + 7.1 - 2.8 + 0.8 + \hat{\varepsilon}_{211} \\
&= 62.6 + \hat{\varepsilon}_{211} \\
\hat{\varepsilon}_{211} &= +1.4
\end{aligned}
$$

- After estimating the size of the mean and main effects $(S_2, \mathcal{O}_1, \text{and } B_1)$ we can then evaluate if they are significant by comparing with the variability in the ε values.

Six Fisher Assumptions for ANOVA

Two assumptions about unknown true values:

$$x_{ijk} = \underbrace{\mu + S_i + \mathcal{O}_j + B_k}_{\text{True Value}} + \underbrace{\varepsilon_{ijk}}_{\text{Error}}$$

1. "C": Unknown true values are <u>constant</u>
2. "A": The components that go together to make x_{ijk} by <u>adding</u> them (not by multiplying etc.)

Four assumptions about errors:

3. "Z": The errors have a mean of zero. eg. $E\{\varepsilon_{ijk}\} = 0$
4. "S": The errors (ε_{ijk}) come from the same distribution, with common standard deviation
5. "I": The errors are independent (eg. there is no relationship between ε_{111} and ε_{112})
6. "N": The distribution of ε_{ijk} is Normal

Assumptions 3-6 can be summarized by

$$\varepsilon_{ijk} \overset{iid}{\sim} N\ (0,\ \sigma^2)$$

"I" "N" "Z" "S": no subscripts on σ^2 implies that σ^2 is the variance, regardless of i, j, or k values

Sums of Squares, df, and Mean Squares

Our goal here is to consider the effects in a model and assess how large each effect is.

In other words, the variability in the response is due to several factors, plus random error.

- How important is each component of variability?
- Which components are significant (ie. large relative to error)?

EX Popcorn Data from one person

		Low Salt	High Salt
Canola	Orville	36	63
	No-name	47	67
Buttery	Orville	28	81
	No-name	42	85

Means		Effect Size
overall:	56.125	—
low salt:	38.25	+/- 17.875
high salt:	74	
canola:	53.25	+/- 2.875
buttery:	59	
no-name:	60.25	+/- 4.125
orville:	52	

Decomposition

Observed Values

36	63
47	67
28	81
42	85

=

Grand Mean

56.125	56.125
56.125	56.125
56.125	56.125
56.125	56.125

+

Salt

-17.875	17.875
-17.875	17.875
-17.875	17.875
-17.875	17.875

+

Oil

-2.875	-2.875
-2.875	-2.875
2.875	2.875
2.875	2.875

+

Pop. Brand

-4.125	-4.125
4.125	4.125
-4.125	-4.125
4.125	4.125

+

Error

4.75	-4
7.5	-8.25
-9	8.25
-3.25	4

We want to quantify the variability of the different factor effects relative to variability errors

Sum of Squares (SS)

The purpose of the SS is to serve as a measure of total variability. We will break the SS_{observed} into pieces as a tool for partitioning the total variability.

- To get SS for observed data, a factor, or residuals:
  square the numbers in the appropriate box and add them up

$$
\begin{aligned}
SS_{\text{observed}} &= 36^2 + 63^2 + \ldots + 85^2 \\
&= 28,297 \longleftarrow \text{measure of total variability}
\end{aligned}
$$

$$
\begin{aligned}
SS_{\text{mean}} &= 56.125^2 + 56.125^2 + \ldots + 56.125^2 \\
&= 8 * 56.125^2 \\
&= 25,200.125 \longleftarrow \text{uninteresting, but helps partition } SS_{\text{observed}}
\end{aligned}
$$

$$
\begin{aligned}
SS_{\text{salt}} &= (-17.875)^2 + 17.875^2 + \dots + 17.875^2 \\
&= 2556.125
\end{aligned}
$$

$$
\begin{aligned}
SS_{\text{oil}} &= (-2.875)^2 + (2.875)^2 + \dots + 2.875^2 \\
&= 66.125
\end{aligned}
$$

$$
\begin{aligned}
SS_{\text{Brand}} &= (-4.125)^2 + (-4.125)^2 + \dots + 4.125^2 \\
&= 136.125
\end{aligned}
$$

$$
\begin{aligned}
SS_{\text{Resid}} &= 4.75^2 + (-4)^2 + \dots + 4^2 \\
&= 338.5
\end{aligned}
$$

Alternative Approach for Calculating SS

$$x_{ijk} = \underbrace{\mu + S_i + \mathcal{O}_j + B_k}_{\text{True Value}} + \underbrace{\varepsilon_{ijk}}_{\text{Error}}$$

where $i = 1, ..., I, j = 1, ..., J, k = 1, ..., K$

2 levels of salt, so I=2 J=2 K=2

$$\begin{aligned}
SS_{\text{Corrected Total}} &= SS_{\text{Total}} - SS_{\text{Mean}} \\
&= \left[\begin{array}{l} \text{sum of squared deviations between each observation's} \\ \text{value and the grand mean} \end{array} \right] \\
&= \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (x_{ijk} - \bar{x})^2 \\
&= (36 - 56.125)^2 + (63 - 56.125)^2 + ... + (85 - 56.125)^2 \\
&= 3096.88 \leftarrow
\end{aligned}$$

often given in computer output instead of SS_{Total} & SS_{Mean}

$$SS_{\text{salt}} = \left[\begin{array}{l} \text{sum of squared deviations between each observation's} \\ \text{salt-group mean and the grand mean} \end{array} \right]$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{x}_{i..} - \bar{x})^2 \qquad \text{where } \bar{x}_{1..} = 74 \text{ and } \bar{x}_{2..} = 38.25$$

$$= (38.25 - 56.125)^2 + (74 - 56.125)^2 + ... + (74 - 56.125)^2$$

1st obs ("36") contribution   from "63"   from "85"

$= 2556.125$

Note: Because the quantity being summed $(\bar{x}_{i..} - \bar{x})^2$ doesn't have j or k subscripts

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{x}_{i..} - \bar{x})^2 = JK\sum_{i=1}^{I}(\bar{x}_{i..} - \bar{x})^2$$

$$= 4(38.25 - 56.125)^2 + 4(74 - 56.125)^2 = 2556.125$$

$$SS_{\text{oil}} = \left[\begin{array}{l} \text{sum of squared deviations between each observation's} \\ \text{oil-group mean and the grand mean} \end{array} \right]$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (\bar{x}_{.j.} - \bar{x})^2 \qquad \text{where } \bar{x}_{.1.} = 53.25 \text{ and } \bar{x}_{.2.} = 59$$

$$= (\underset{\text{from "36"}}{53.25} - 56.125)^2 + (\underset{\text{from "63"}}{53.25} - 56.125)^2 + \ldots + (\underset{\text{from "85"}}{59} - 56.125)^2$$

$$= IK \sum_{j=1}^{J} (\bar{x}_{.j.} - \bar{x})^2$$

$$= 66.125$$

$$SS_{\text{brand}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (\bar{x}_{..k} - \bar{x})^2$$

$$= IJ \sum_{k=1}^{K} (\bar{x}_{..k} - \bar{x})^2$$

$$= 136.125$$

ANOVA Table

Source	df	SS	MS	F	p-value
Mean		25,200.125			
Salt		2,556.125			
Oil		66.125			
Brand		136.125			
Residual Error		338.5			
TOTAL		28,297			

- But comparing SS_{brand} to SS_{Resid} isn't really fair, because more "units of information" are used to calculate the SS_{Resid}.
  - degrees of freedom allows us to make comparisons

Degrees of Freedom

Begin by thinking of your total number of measurements (N). This is the number of informational units you have to "spend" on building your model.

$$x_{ijk} = \underbrace{\mu + S_i + \mathcal{O}_j + B_k}_{\text{True Value}} + \underbrace{\varepsilon_{ijk}}_{\text{Error}}$$

- Strategy: Estimate the components of the "true value" part of the model, with as many informational units as possible saved to estimate the "error" part of the model.

Then,

df= units of information needed to describe the factor

- The df associated with each factor (or error) will be a different number.
- Sum of the df for a model equals the total number of measurements

Calculating df

Remember the decomposition :

Observed Values

36	63
47	67
28	81
42	85

=

Grand Mean

56.125	56.125
56.125	56.125
56.125	56.125
56.125	56.125

+

Salt

-17.875	17.875
-17.875	17.875
-17.875	17.875
-17.875	17.875

+

Oil

-2.875	-2.875
-2.875	-2.875
2.875	2.875
2.875	2.875

+

Pop. Brand

-4.125	-4.125
4.125	4.125
-4.125	-4.125
4.125	4.125

+

Error

4.75	-4
7.5	-8.25
-9	8.25
-3.25	4

df for a factor is the number of **free** numbers in the associated decomposition table – the number of slots you can freely fill in, after accounting for the pattern of repetitions and zero-sum properties

Observed Values

36	63
47	67
28	81
42	85

df=____

=

Grand Mean

56.125	56.125
56.125	56.125
56.125	56.125
56.125	56.125

df=____

+

Salt

-17.875	17.875
-17.875	17.875
-17.875	17.875
-17.875	17.875

df=____

+

Oil

-2.875	-2.875
-2.875	-2.875
2.875	2.875
2.875	2.875

df=____

+

Pop. Brand

-4.125	-4.125
4.125	4.125
-4.125	-4.125
4.125	4.125

df=____

+

Error

4.75	-4
7.5	-8.25
-9	8.25
-3.25	4

df=____

Mean Square (MS)

MS is an average variability for a factor

$$MS = \frac{SS}{df}$$

$$\text{eg. } MS_{salt} = \frac{SS_{salt}}{df_{salt}} = \frac{2556.125}{1} = 2556.125$$

Now, these numbers can be fairly compared

$$\text{eg. } MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{338.5}{4} = 84.625$$

ANOVA Table

Source	df	SS	MS	F	p-value
Mean	1	25,200.125	25,200.125		
Salt	1	2,556.125	2,556.125		
Oil	1	66.125	66.125		
Brand	1	136.125	136.125		
Residual Error	4	338.5	84.625		
TOTAL	8	28,297			

Standard Deviation = MS_{Error} (aka "MS_{Resid}" or "MSE")

The value $MS_{Error} = 84.625$ has a particularly important meaning

- $\sqrt{MSE} = 9.199$ is the SD for this analysis. This is the estimate of σ_ε, the standard deviation of the ε's
  - Note that this SD is <u>different</u> from the sample standard deviation of the eight observations $(36, 63, \cdots, 85)$, which is calculated:

$$\sqrt{\frac{1}{8-1}\left[(36-56.125)^2 + \cdots + (85-56.125)^2\right]}$$

  - For clarity, I'll call it "$\sqrt{MSE}$" or "$\hat{\sigma}_\varepsilon$"

- $\hat{\sigma}_\varepsilon$ informs about the typical size of an ε

  observed value= true value $\pm\hat{\sigma}_\varepsilon$
- If $\hat{\varepsilon}$ values are normally distributed,
  - 68% of $\hat{\varepsilon}$ values are in $(-\hat{\sigma}_\varepsilon, \hat{\sigma}_\varepsilon)$
  - 95% of $\hat{\varepsilon}$ values are in $(-2\hat{\sigma}_\varepsilon, 2\hat{\sigma}_\varepsilon)$
  - 99.7% of $\hat{\varepsilon}$ values are in $(-3\hat{\sigma}_\varepsilon, 3\hat{\sigma}_\varepsilon)$

You can use the $\hat{\varepsilon}$ values $(4.75, -4, \cdots, 4)$ to check assumptions: CAZSIN

- Z: Always true for $\hat{\varepsilon}$ values
- S: If there's replication within treatment groups, compare the standard deviation of responses (or $\hat{\varepsilon}$ values) WITHIN each treatment
- I: Plot $\hat{\varepsilon}$ values in order of measurements; check for trend
- N: Look at histogram for $\hat{\varepsilon}$ values

ANOVA Hypothesis Tests

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$
vs.
H_a: At least one μ_i is different from the others

In other words...

H_0: Observed differences in group means are due to chance

e.g. salt-group

H_a: Observed differences are <u>real</u>

To test H_0, we will use the F-ratio

$$\text{F-ratio} = \frac{MS_{group}}{MS_{Error}} = \frac{\text{Variability due to group}}{\text{Variability due to error}}$$

$$\text{F-ratio} = \frac{\text{MS group}}{\text{MS Error}} \longleftarrow \text{ eg. } MS_{salt} \text{ or } MS_{oil}$$

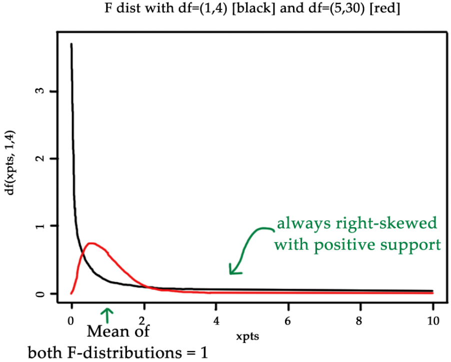
- MS_{Error} quantifies variability due to error
- MS_{group} is based on estimated group means that are <u>also</u> subject to error (remember if x has variance σ^2, then $\bar{x}$ isn't a perfect estimate of μ —it will have variance $\frac{\sigma^2}{n}$)
  - So, MS_{group} is partly a measure of variability due to group, and partly a measure of variability due to error
  - df chosen to be the correct divisor of the SS so that the measure of variability due to error in the MS_{group} is comparable to the measure of variability due to error in MS_{Error}

$$\text{F-ratio} = \frac{MS_{group}}{MS_{Error}} = \frac{\begin{array}{c}\text{One estimate of}\\ \text{chance error}\\ \text{variability}\end{array} + \begin{array}{c}\text{Estimate of}\\ \text{variability due}\\ \text{to groups}\end{array}}{\begin{array}{c}\text{Another estimate}\\ \text{of chance error}\\ \text{variability}\end{array}}$$

- Under H_0 (no group differences), the numerator and denominator estimate the same number and the F-ratio will be near 1.
- If group differences are large, F will be much greater than 1.

Under H_o, the ratio of similar estimates of variability will follow the F-distribution with degrees of freedom df_{group} and df_{Error}, or:

$$F = \frac{MS_{group}}{MS_{Error}} \sim \mathrm{F}_{\mathrm{df}_{group},\mathrm{df}_{error}}$$



How extreme does the F-statistic need to be before we consider it to be significant?

$$\text{p-value} = \Pr\{\mathrm{F}_{df_1,df_2} > \underset{\substack{\uparrow \\ \text{calculated}}}{\mathrm{F}} \mid H_0 \text{ is true}\}$$

- Find this with "`1 - pf(Fstat, df1, df2)`" in R

Significant F-statistics should indicate significant differences but...

- may indicate an inadequate model
- may indicate assumptions are not met (CASZIN)
- will not indicate if the results are practically significant

ANOVA Table

Source	df	SS	MS	F	p-value
Mean	1	25,200.125	25,200.125		
Salt	1	2,556.125	2,556.125	30.207	0.0053
Oil	1	66.125	66.125	0.781	0.4266
Brand	1	136.125	136.125	1.609	0.2734
Residual Error	4	338.5	84.625		
TOTAL	8	28,297			

Confidence Intervals

C.I. for differences in means:

$$\underbrace{\underbrace{\bar{x}_1 - \bar{x}_2}_{\substack{\text{estimate of}\\ \mu_1 - \mu_2}} \pm \underbrace{t_{\alpha/2,\nu}}_{\substack{\text{table value}\\ \nu = \text{df}_{error}}} \underbrace{\sqrt{\underbrace{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}_{\substack{\text{\# of obs}\\ \text{used in each mean}}} \underbrace{\text{MS}_{Error}}_{\substack{\text{from ANOVA}\\ \text{table}}}}}_{\text{Standard Error}}}_{\text{Margin of Error}}$$

- Note that when the degrees of freedom ν drops below roughly 5, the $t_{\alpha/2}$ value increases sharply
  - It's almost always worth it to add more experimental units when $\nu \leq 5$.