One factor completely randomized design

Example: 12 mice randomly assigned to 3 diets, with 4 mice to each diet. Randomly select 4 mice out of 12 and assign them to diet 1, randomly select 4 out of the remaining 8 and assign them to diet 2 and assigning the last 4 mice to diet 3. This is a **completely randomized design**. Lifelength (in month) recorded below.
Factor: Type of diet. Treatments: 3 diets. Response: Lifelength.

diet 1 (high calorie) 22 18 21 22
diet 2(medium calorie) 20 19 23 21
diet 3(low calorie) 23 24 20 25

Assume $y_{11}, y_{12}, y_{13}, \cdots, y_{1n_1} \sim N(\mu_1, \sigma^2)$,
$y_{21}, y_{22}, y_{23}, \cdots, y_{2n_2} \sim N(\mu_2, \sigma^2)$,
$y_{31}, y_{32}, y_{33}, \cdots, y_{3n_3} \sim N(\mu_3, \sigma^2)$.
y_{ij} : j th observation on treatment i.

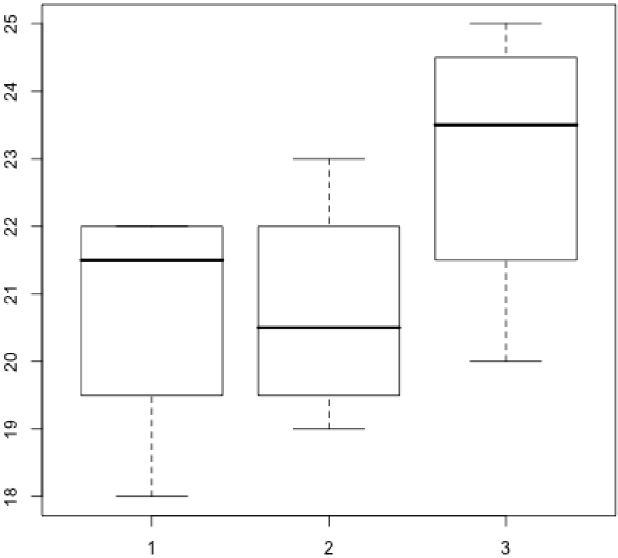
Assumptions

General assumptions:

- random, independent samples. Independent observations within each sample.
- Each sample comes from a normal distribution.
- Equal population variance $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma^2$.

Goal: test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$.
H_a : not all μ_i's are equal.

boxplot of data



Treatment effect

Suppose there are t treatment groups:
Write $\mu_i = \mu + \alpha_i$, where $\mu = \frac{\sum \mu_i}{t}$ is grand population mean and α_i indicates deviation of treatment group mean from grand mean. α_i is called ***i*th treatment effect.**

e.g., $\mu_1 = 10, \mu_2 = 12, \mu_3 = 23$, then $\mu = 15$,
Write $\mu_1 = \mu + \alpha_1 = 15 - 5$, so $\alpha_1 = -5$
$\mu_2 = \mu + \alpha_2 = 15 - 3$, so $\alpha_2 = -3$
$\mu_3 = \mu + \alpha_3 = 15 + 8$, so $\alpha_3 = 8$.
Note $\sum_{i=1}^{3} \alpha_i = 0$.
If μ_i's are very different, then α_i' s will be big in magnitude.
If μ_i's are equal then all α_i's are 0.
Alternative way to write $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_t = 0$.
Estimate of α_i : $\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$ where $\bar{y}_{\cdot\cdot} = \frac{\sum_i \sum_j y_{ij}}{N}$ is sample grand mean and $N = \sum_{i=1}^{t} n_i$.

ANOVA

ANOVA measures the variability among groups means (i.e., treatment effect size) relative to the experimental error.

Variation BETWEEN groups: sum of squared difference between group mean and grand mean.
SSTR$= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{t} n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{t} n_i \hat{\alpha}_i^2$
Here n_i is sample size in group i,
SSTR: Treatment sum of squares.
Steps to get SSTR:

- Step 1: In each treatment group, compute $(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$ then multiply by n_i.
- Step 2: Add across the treatment groups.

Treatment mean square= MSTR=SSTR/d.f.=SSTR/(t-1)

MSE

Variation WITHIN groups: sum of squared difference between observed value and group mean. Measures experimental error.
$y_{ij} - \bar{y}_{i\cdot}$ is also called **error** or **residual.**
SSE=$\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$
SSE: sum of squared errors (or residuals)

Mean squared error =MSE= SSE/d.f.=SSE/(N-t)
MSE is just S_p^2 in two sample case.
Fact: E(MSE)=σ^2 i.e., MSE is an unbiased estimator for σ^2.
Note $MSE = \frac{n_1-1}{N-t} s_1^2 + \frac{n_2-1}{N-t} s_2^2 + \cdots + \frac{n_t-1}{N-t} s_t^2$ which is weighted average of the sample variances. When $n_1 = n_2 = \cdots = n_t$, MSE is the simple average of the sample variances.

MSTR/MSE

MSTR measures variations due to both treatment effects and experimental error.
MSE measures variations due to experimental error only.
E(MSTR)$= \sigma^2 + \frac{1}{t-1}\sum_{i=1}^{t} n_i\alpha_i^2$
E(MSE)$=\sigma^2$.
under H_0, MSTR/MSE should be close 1.
If the ratio $\frac{MSTR}{MSE}$ is much larger than 1, it indicates treatment effects.

Example

Summary of data: $\bar{y}_{1.} = 20.75, s_1 = 1.89; \bar{y}_{2.} = 20.75, s_2 = 1.71; \bar{y}_{3.} = 23.00, s_3 = 2.16; \bar{y}_{..} = 21.50.$

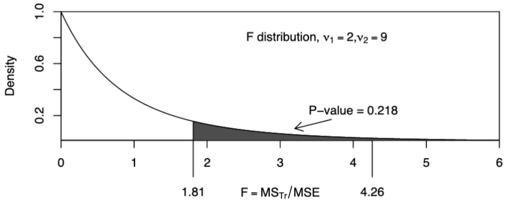
SSTR$=4*(20.75-21.50)^2+4*(20.75-21.50)^2+4*(23-21.50)^2 = 13.50.$
MSTR$=13.50/(3-1) = 6.75.$
MSE$=\frac{1}{3}(s_1^2 + s_2^2 + s_3^2) = 3.72.$ (equal sample sizes)

The ratio is $\frac{MSTR}{MSE} = 1.81.$

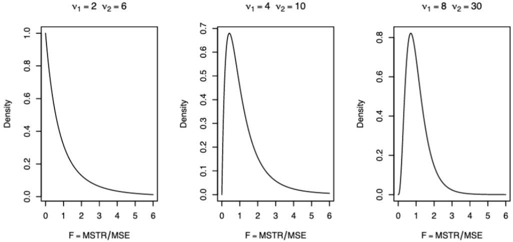
The F distribution

If the t independent random samples are from t normal distributions with identical population variances, and $H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$ is true, then $\frac{MSTR}{MSE}$ has a F distribution with numerator d.f. t-1 and denominitor d.f. N-t, where $N = \sum_i n_i$.

In the example, $F = MSTR/MSE \sim F_{2,9}$.
The p-value $= P(F > 1.81) = 0.22$. (Large F values support H_a, so only look at right tail probability)
R code: 1-pf(1.81,2,9).
or pf(1.81,2,9,lower.tail=F)



F distribution



F distribution is right skewed and a random variable with a F distribution takes non-negative values.

exercise

An engineered varied the percentage of steel added to metal tips and the strength was measured for each tip.

percent	strength
15 percent :	7 11 7 15 9
20 percent :	12 18 17 12 18
25 percent :	14 18 19 19 18
30 percent :	19 25 22 19 23

Compute MSTR and MSE.

Compute MSTR and MSE

```
SSTR=363.4,
MSTR=121.13
MSE=8.025
f=MSTR/MSE=15.1
```

```
p-value < 0.0001
```

```
R: 1-pf(15.1,3,16)
```

Decompose SST

SSTR: treatment sum of squares
$SSTR = \sum_{i=1}^{t} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$

SSE: sum of squared errors.
$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2$

SST: total sum of squares (corrected for mean):
$SST = \sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2$ (measure total variation in data)

Fact: $SST = SSTR + SSE$.

ANOVA table

ANOVA table: one factor

Source of Variation	Df	SS	MS	F	p-value
Treatment	t-1	SSTR	MSTR	MSTR/MSE	
Error	N-t	SSE	MSE		
Total(corrected)	N-1	SST			

Note $N-1=(t-1)+(N-t)$ and SST= SSTR+SSE

R code

```
g1 = c(7, 7, 15, 11, 9)
g2 = c(12, 17, 12, 18, 18)
g3 = c(14, 19, 19, 18, 18)
g4 = c(19, 25, 22, 19, 23)
y =c(g1,g2,g3,g4)
steelcontent = c(rep(1,5),rep(2,5),rep(3,5),rep(4,5))
steelcontent = factor (steelcontent)
output = aov(y~steelcontent)
summary(output)
```

steelcontent is a categorical variables and 1, 2, 3, 4 are just labels for its levels

Data format

	y	steelcontent
1	7	1
2	7	1
3	15	1
4	11	1
5	9	1
6	12	2
7	17	2
8	12	2
9	18	2
10	18	2
11	14	3
12	19	3
13	19	3
14	18	3
15	18	3
16	19	4
17	25	4

R ANOVA table

```
> summary(output)
Analysis of Variance Table

Response: y
             Df  Sum Sq   Mean Sq   F value     Pr(>F)
steelcontent  3  363.40   121.13     15.095    6.315e-05 **
Residuals    16  128.40     8.03
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Uncommon ANOVA table

ANOVA table: one factor, uncorrected for the mean

Source of Variation	Df	SS	MS	F	p-value
Grand Mean	1	SSGM	MSGM		
Treatment	t-1	SSTR	MSTR	MSTR/MSE	
Error	N-t	SSE	MSE		
Total(uncorrected)	N	SST			

SSGM=$N(\bar{y}_{\cdot\cdot}^2)$ and
$SST_{uncorrected} = \sum_i \sum_j y_{ij}^2$
Note $SST_{uncorrected} = SST_{corrected} + SSGM$,
now we have $SST_{uncorrected} = SSGM + SSTR + SSE$, and
$N = 1 + (t-1) + (N-t)$

t test and F test

Two independent sample pooled t test for testing
$H_a : \mu_1 \neq \mu_2$(assuming equal variance) is equivalent to one way ANOVA F test with $t = 2$.
In fact $t^2 = F$.

example

Researchers are interested in comparing the effects of two programs to enhance short term memory. Five students were randomly selected from 10 students and assigned to program A. The remaining 5 students were assigned to program B. The number of words out of 35 words each student could remember was recorded.
A: 18,24,30,21,32
B: 22,29,34,25,35
$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$.
Summary of data: $\bar{y}_{1\cdot} = 25, s_1 = 5.92, \bar{y}_{2\cdot} = 29, s_2 = 5.61$.

Compute s_p^2 and the test statistic t. The p-value is (   ).
Compute SSTR and MSTR. The F-value is (   ). The p-value is (   ).

Exercise: unequal sample sizes

Rats were given one of four different diets at random, and the response measure was liver weight as a percentage of body weight. The responses were

Treatment			
1	2	3	4
3.52	3.47	3.54	3.74
3.36	3.73	3.52	3.83
3.57	3.38	3.61	3.87
4.19	3.87	3.76	4.08
3.88	3.69	3.65	4.31
3.76	3.51	3.51	3.98
3.94	3.35		3.86
	3.64		3.71

(a) Compute the overall mean and treatment effects.

(b) Compute the Analysis of Variance table for these data. What would you conclude about the four diets?

Data format

	y	diet
1	3.52	1
2	3.36	1
3	3.57	1
4	4.19	1
5	3.88	1
6	3.76	1
7	3.94	1
8	3.47	2
9	3.73	2
10	3.38	2
11	3.87	2
12	3.69	2
13	3.51	2
14	3.35	2
15	3.64	2
16	3.54	3
17	3.52	3

18	3.61	3
19	3.76	3
20	3.65	3
21	3.51	3
22	3.74	4
23	3.83	4
24	3.87	4
25	4.08	4
26	4.31	4
27	3.98	4
28	3.86	4
29	3.71	4

Solutions

```
> g1=c(3.52,3.36,3.57,4.19,3.88,3.76,3.94)
> g2=c(3.47,3.73,3.38,3.87,3.69,3.51,3.35,3.64)
> g3=c(3.54,3.52,3.61,3.76,3.65,3.51)
> g4=c(3.74,3.83,3.87,4.08,4.31,3.98,3.86,3.71)
> y=c(g1,g2,g3,g4)
> diet=c(rep(1,7),rep(2,8),rep(3,6),rep(4,8))
> diet=factor(diet)
> output=aov(y~diet)
> summary(output)
            Df Sum Sq Mean Sq F value Pr(>F)
diet         3 0.5782 0.19274   4.658 0.0102 *
Residuals   25 1.0344 0.04138
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
> boxplot(y~diet)
```

Solutions

```
> m1=mean(g1);m2=mean(g2);m3=mean(g3);m4=mean(g4)
> grandm=mean(y)
> grandm
[1] 3.718276
> SSTR=7*(m1-grandm)^2+8*(m2-grandm)^2
    +6*(m3-grandm)^2+8*(m4-grandm)^2
> SSTR
[1] 0.578209
> v1=var(g1);v2=var(g2);v3=var(g3);v4=var(g4)
> MSE=6/25*v1+7/25*v2+5/25*v3+7/25*v4
> MSE
[1] 0.04137619
```

Exercise: 4.4. Melting time of 15 ice cubes in 3 types of beverages:

Coca Cola	19, 17, 15, 14, 18
Orange Juice	27, 28, 30, 26, 27
Water	10, 11, 13, 7, 9

What is the factor of interest?
What are the treatments?
What are the experimental units?
Is there evidence of melting time is the same for the three treatments? Use an F test to answer this question at 0.05 level of significance.

Problem 4.6

Treatment Group	n	Mean	Standard Deviation
Group Exercise	24	4.5	8.4
Physiotherapy	35	4.1	8.0
Osteopathy	39	5.0	10.5

What is SSE? What is MSE in the ANOVA table?