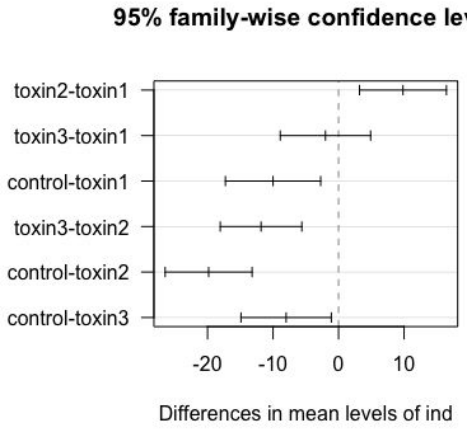
Two-way ANOVA (univariate)

ANOVA table

07/29/2021

In the previous lecture,



- Simultaneous pairwise comparisons:
  - Bonferroni method: adjusting with $1-\alpha/m$.
  - Tukey's method: studentized range distribution.

$$\text{Tukey's CI} \subseteq \text{Bonferroni CI}.$$

- Non-parametric Kruskal-Wallis test:
  - It extends Mann-Whitney test and focuses on medians;
  - The test statistic is:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\frac{R_i}{n_i} - \frac{n+1}{2} \right)^2.$$

  - The rejection region is $\{H \geq c\}$.

- Two-way factorial design:
  - Treatment means plot: check non-parallel lines to see whether there are interaction effects.
  - Model assumption:

$$Y_{ijl} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijl}, \quad \epsilon_{ijl} \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Two-way ANOVA: LRT

Proposition D. Assume $Y_{ijl} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijl}$, where $\epsilon_{ijl} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. We can derive the maximum likelihood in $\Theta = \Theta_0 \cup \Theta_1$:

$$\sup_{\Theta} L = \left(\frac{1}{\sqrt{2\pi\hat{\sigma}_n^2}} \right)^n e^{-\frac{n}{2}},$$

where $\hat{\sigma}_n^2 = n^{-1} \sum_i \sum_j \sum_{l=1}^{n_{ij}} \left(Y_{ijl} - \bar{Y}_{ij\cdot} \right)^2$.

Sum of squares within cells:

$$SS_E = \sum_{i=1}^{I} \sum_{j}^{J} \sum_{l=1}^{n_{ij}} \left(Y_{ijl} - \bar{Y}_{ij\cdot} \right)^2.$$

Corollary D. Under the two-way ANOVA assumptions,

$$\frac{SS_E}{\sigma^2} \sim \chi^2_{n-IJ}, \text{ and it's independent of each cell mean } \bar{Y}_{ij\cdot},$$

where $i = 1, \ldots, I, j = 1, \ldots, J$.

Two-way ANOVA: LRT for factor A

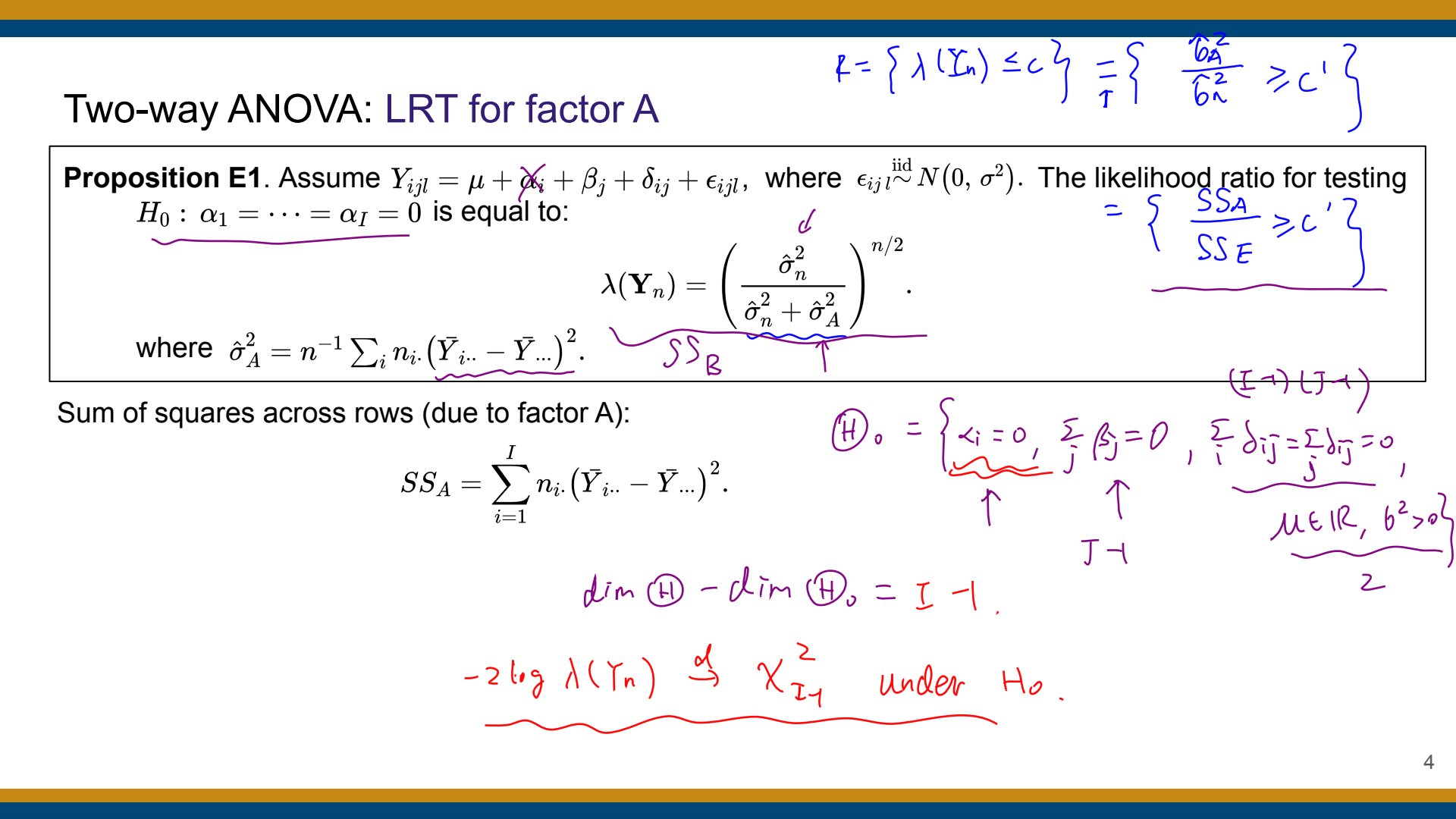
Proposition E1. Assume $Y_{ijl} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijl}$, where $\epsilon_{ijl} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The likelihood ratio for testing $H_0 : \alpha_1 = \cdots = \alpha_I = 0$ is equal to:

$$\lambda(\mathbf{Y}_n) = \left(\frac{\hat{\sigma}_n^2}{\hat{\sigma}_n^2 + \hat{\sigma}_A^2} \right)^{n/2}.$$

where $\hat{\sigma}_A^2 = n^{-1} \sum_i n_{i\cdot} \left(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots} \right)^2$.

Sum of squares across rows (due to factor A):

$$SS_A = \sum_{i=1}^{I} n_{i\cdot} \left(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots} \right)^2.$$



Proof. We want to calculate $\sup_{H_0} L$ under $H_0: \alpha_1 = \cdots = \alpha_I = 0$.

$$\sup_{H_0} \ell(\vec{\alpha}, \vec{\beta}, \vec{\delta}, \mu, \sigma^2) = \sup_{H_0}\left[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i\sum_j\sum_{l=1}^{n_{ij}} (Y_{ijl} - \mu - \beta_j - \delta_{ij})^2 \right]$$

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_i\sum_j\sum_{l=1}^{n_{ij}} (Y_{ijl} - \mu - \beta_j - \delta_{ij}) = 0$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J-1}\sum_{l=1}^{n_{ij}} (Y_{ijl} - \mu - \beta_j - \delta_{ij})^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{l=1}^{n_{iJ}} (Y_{iJl} - \mu - \beta_J - \delta_{iJ})^2$$

$$\frac{\partial \ell}{\partial \beta_j} = \frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{l=1}^{n_{ij}} (Y_{ijl} - \mu - \beta_j - \delta_{ij}) - \frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{l=1}^{n_{iJ}} (Y_{iJl} - \mu - \beta_J - \delta_{iJ}) = 0$$

$\beta_J = \sum_{j=1}^{J-1} \beta_j$

$j = 1, \cdots, J-1$

Sum all the likelihood equations for $\frac{\partial \ell}{\partial \beta_j} = 0$:

$$\frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} (Y_{ijl} - \mu - \beta_j - \delta_{ij}) - \frac{J}{\sigma^2}\sum_{i=1}^{I}\sum_{l=1}^{n_{iJ}} (Y_{iJl} - \mu - \beta_J - \delta_{iJ}) = 0$$

$$= 0$$

$$\sum_{i=1}^{I}\sum_{l=1}^{n_{ij}} (Y_{ijl} - \mu - \beta_j - \delta_{ij}) = 0$$

Assume balance design: $n_{ij} = n$

$$Y_{\cdot j \cdot} - n\sum_{i=1}^{I}(\mu + \beta_j) - n\sum_{i=1}^{I}\delta_{ij} = 0 \Rightarrow \mu + \beta_j = \bar{Y}_{\cdot j \cdot}$$

$$\frac{\partial l}{\partial \delta_{ij}} = \frac{1}{\sigma^2}\sum_{l=1}^{n_{ij}}(Y_{ijl} - \mu - \beta_j - \delta_{ij}) + \frac{1}{\sigma^2}\sum_{l=1}^{n_{IJ}}(Y_{IJl} - \mu - \beta_J - \delta_{IJ})$$

$$\downarrow \quad -\frac{1}{\sigma^2}\sum_{l=1}^{n_{ij}}(Y_{IJl} - \mu - \beta_j - \delta_{Ij}) - \frac{1}{\sigma^2}\sum_{l=1}^{n_{iJ}}(Y_{iJl} - \mu - \beta_J - \delta_{iJ})$$

$i = 1, \cdots, I-1$

$J = 1, \cdots, J-1$

Sum all $\frac{\partial l}{\partial \delta_{ij}} = 0$ to get:

$$0 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} + (I-1)(J-1)\sum_{l=1}^{n_{IJ}} - (I-1)\sum_{j=1}^{J}\sum_{l=1}^{n_{Ij}} - (J-1)\sum_{i=1}^{I}\sum_{l=1}^{n_{iJ}}$$

$IJ - I - J$

$$0 = IJ\sum_{l=1}^{n_{IJ}}(Y_{IJl} - \mu - \beta_J - \delta_{IJ}) - I\sum_{j=1}^{J}\sum_{l=1}^{n_{Ij}}(Y_{Ijl} - \mu - \beta_j - \delta_{Ij})$$

$$IM\bar{Y}_{j\cdot} = Y_{\cdot j\cdot}$$

$$\frac{1}{IM}\left(Y_{IJ\cdot} - M\bar{Y}_{\cdot J\cdot} - M\delta_{IJ}\right) - \left(\frac{Y_{I\cdots}}{JM} - \frac{Y_{\cdots}}{IJM}\right) = 0$$

$$\bar{Y}_{IJ\cdot} - \bar{Y}_{\cdot J\cdot} - \delta_{IJ} - \bar{Y}_{I\cdots} + \bar{Y}_{\cdots} = 0 \Rightarrow \boxed{\delta_{IJ} = \bar{Y}_{IJ\cdot} - \bar{Y}_{\cdot J\cdot} - \bar{Y}_{I\cdots} + \bar{Y}_{\cdots}}$$

$$\mu+\rho_J+\delta_{IJ} = \cancel{\bar{Y}_{\cdot J\cdot}} + \bar{Y}_{IJ\cdot} \cancel{-\bar{Y}_{\cdot J\cdot}} - \bar{Y}_{I\cdot\cdot} + \bar{Y}_{\cdots}$$

$$= \bar{Y}_{IJ\cdot} - \bar{Y}_{I\cdot\cdot} + \bar{Y}_{\cdots}$$

$$\Rightarrow \mu+\rho_j+\delta_{ij} = \bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} + \bar{Y}_{\cdots}$$

$$\Rightarrow \sup_{(H)} L = \left(\frac{1}{\sqrt{2\pi\hat{\sigma}_0^2}}\right)^n e^{-\frac{n}{2}} \quad \text{where } \hat{\sigma}_0^2 = \frac{1}{n}\sum_i\sum_j\sum_{l=1}^{n_{ij}}\left(Y_{ijl} - \bar{Y}_{ij\cdot} + \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots}\right)^2$$

$$= \frac{1}{n}\sum_i\sum_j\sum_{l=1}^{n_{ij}}\left[\left(Y_{ijl}-\bar{Y}_{ij\cdot}\right)^2 + \left(\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdots}\right)^2 \cancel{+2\left(Y_{ijl}-\bar{Y}_{ij\cdot}\right)\left(\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdots}\right)}\right]$$

$$\left(\text{cross term} = 0\right)$$

$$= \frac{1}{n}\sum_i\sum_j\sum_{l=1}^{n_{ij}}\left(Y_{ijl}-\bar{Y}_{ij\cdot}\right)^2 \quad \hat{\sigma}_n^2$$

$$+ \frac{1}{n}\sum_i n_{i\cdot}\left(\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdots}\right)^2 \quad \hat{\sigma}_A^2$$

$$\Rightarrow \lambda(Y_n) = \left(\frac{\hat{\sigma}_n^2}{\hat{\sigma}_0^2}\right)^{\frac{n}{2}}$$

$$= \left(\frac{\hat{\sigma}_n^2}{\hat{\sigma}_A^2+\hat{\sigma}_n^2}\right)^{\frac{n}{2}} \blacksquare$$

Two-way ANOVA: LRT for factor B

$R = \left\{ \frac{SS_B}{SS_E} \geq c' \right\}$

Proposition E2. Assume $Y_{ijl} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijl}$, where $\epsilon_{ijl} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The likelihood ratio for testing $H_0 : \beta_1 = \cdots = \beta_J = 0$ is equal to:

$$\lambda(\mathbf{Y}_n) = \left(\frac{\hat{\sigma}_n^2}{\hat{\sigma}_n^2 + \hat{\sigma}_B^2} \right)^{n/2}.$$

where $\hat{\sigma}_B^2 = n^{-1} \sum_j n_{\cdot j} \left(\bar{Y}_{\cdot j \cdot} - \bar{Y}_{\cdots} \right)^2$.

Sum of squares across columns (due to factor B):

$$SS_B = \sum_{j=1}^{J} n_{\cdot j} \left(\bar{Y}_{\cdot j \cdot} - \bar{Y}_{\cdots} \right)^2.$$

$\dim \Theta - \dim \Theta_0 = J - 1$

Two-way ANOVA: LRT for A*B

Proposition E3. Assume $Y_{ijl} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijl}$, where $\epsilon_{ijl} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The likelihood ratio for testing H_0 : all δ_{ij}'s are zero is equal to:

$$\lambda(\mathbf{Y}_n) = \left(\frac{\hat{\sigma}_n^2}{\hat{\sigma}_n^2 + \hat{\sigma}_{AB}^2} \right)^{n/2}.$$

where $\hat{\sigma}_{AB}^2 = \sum_i \sum_j n_{ij} \left(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots} \right)^2$.

Sum of squares among cells (due to interaction between A and B):

$$SS_{AB} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \left(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots} \right)^2.$$

Two-way ANOVA: Summary

Theorem F. Under the assumptions of two-way ANOVA model,

1) Denote the total sum of squares as

$$SS_{\text{Tot}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} \left(Y_{ijk} - \bar{Y}_{\cdots}\right)^2.$$

Then $SS_{\text{Tot}} = SS_A + SS_B + SS_{AB} + SS_E$.

2) Under $H_0 : \alpha_1 = \cdots = \alpha_I = 0$,

$$\frac{SS_A}{\sigma^2} \sim \chi^2_{I-1}.$$

3) Under $H_0 : \beta_1 = \cdots = \beta_J = 0$,

$$\frac{SS_B}{\sigma^2} \sim \chi^2_{J-1}.$$

4) Under $H_0 :$ all δ_{ij}'s are zero,

$$\frac{SS_{AB}}{\sigma^2} \sim \chi^2_{(I-1)(J-1)}.$$

5) SS_A, SS_B, SS_{AB} and SS_E are mutually independent.

$\sum_i n_{i\cdot}(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots})^2$ ← a function of $\bar{Y}_{ij\cdot}$.

Two-way ANOVA: Summary

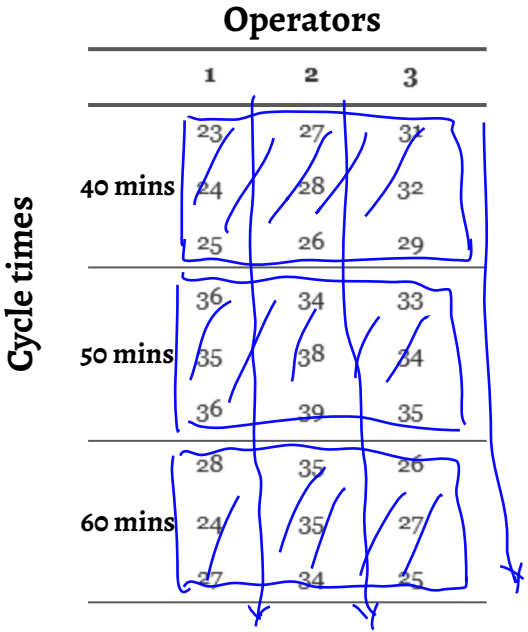
$R = \left\{ \frac{SS_A}{SS_E} \geq c' \right\}$

$= \left\{ \frac{SS_A/(I-1)}{SS_E/(n-IJ)} \geq c'' \right\}$

Source	df	Sum Sq	Mean Sq	*F* value	*p*-value
Factor A	$I-1$	$SS_A = \sum_{i=1}^{I} n_{i\cdot}(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots})^2$	$MS_A = SS_A/(I-1)$	MS_A / MS_E	`pf(MS_A/MS_E, I-1, n-I*J, lower.tail = F)`
Factor B	$J-1$	$SS_B = \sum_{j=1}^{J} n_{\cdot j}(\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots})^2$	$MS_B = SS_B/(J-1)$	MS_B / MS_E	`pf(MS_B/MS_E, J-1, n-I*J, lower.tail = F)`
Interaction A*B	$(I-1)(J-1)$	$SS_{AB} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots})^2$	$MS_{AB} = SS_{AB}/\{(I-1)(J-1)\}$	MS_{AB} / MS_E	`pf(MS_AB/MS_E, (I-1)*(J-1), n-I*J, lower.tail = F)`
Error	$n-IJ$	$SS_E = \sum_{i=1}^{I}\sum_{j}^{J}\sum_{l=1}^{n_{ij}} (Y_{ijl} - \bar{Y}_{ij\cdot})^2$	$MS_E = SS_E/(n-IJ)$		
Total	$n-1$	$SS_{\text{Tot}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{\cdots})^2$			

Two treatment factors

Example 4 *cont'd*. A fabric plant is studying the effect of several factors on the dyeing of cotton-synthetic cloth. At 300°C, three operators and three cycle times were selected. The finished cloth was compared to a standard, and a numerical score was assigned.



$$SS_A = \sum_{i=1}^{I} n_{i\cdot}\left(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots}\right)^2$$

cycle times operator obs

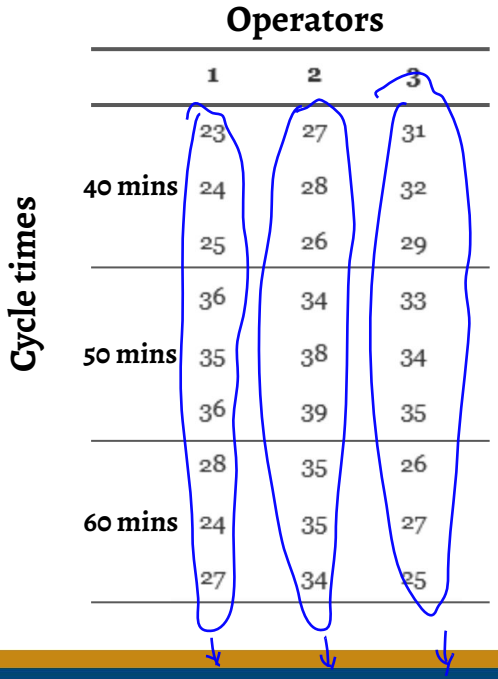
```
> ## Bookkeeping
> I = 3; J = 3; n = nrow(Obs_w_names)
> all_mean = mean(Obs_w_names$Obs)

## Calculate SS_A
> S_A = aggregate(Obs ~ cycle_times, data = Obs_w_names,
          FUN = function(vec) length(vec)*(mean(vec)-all_mean)^2)
> S_A
  cycle_times       Obs
1          40 102.23457
2          50 221.67901
3          60  22.82716

> SS_A = sum(S_A$Obs)
> SS_A
[1] 346.7407
```

Two treatment factors

Example 4 *cont'd*. A fabric plant is studying the effect of several factors on the dyeing of cotton-synthetic cloth. At 300°C, three operators and three cycle times were selected. The finished cloth was compared to a standard, and a numerical score was assigned.



Cycle times	Operators 1	2	3
40 mins	23	27	31
	24	28	32
	25	26	29
50 mins	36	34	33
	35	38	34
	36	39	35
60 mins	28	35	26
	24	35	27
	27	34	25

$$SS_B = \sum_{j=1}^{J} n_{\cdot j} \left(\bar{Y}_{\cdot j \cdot} - \bar{Y}_{\dots}\right)^2$$

```
## Calculate SS_B
> S_B = aggregate(Obs ~ operator, data = Obs_w_names,
          FUN = function(vec) length(vec)*(mean(vec)-all_mean)^2)

> SS_B = sum(S_B$Obs)
> SS_B
[1] 82.07407
```

Two treatment factors

Example 4 *cont'd*. A fabric plant is studying the effect of several factors on the dyeing of cotton-synthetic cloth. At 300°C, three operators and three cycle times were selected. The finished cloth was compared to a standard, and a numerical score was assigned.

Cycle times	Operators 1	2	3
40 mins	23	27	31
	24	28	32
	25	26	29
50 mins	36	34	33
	35	38	34
	36	39	35
60 mins	28	35	26
	24	35	27
	27	34	25

$$SS_E = \sum_{i=1}^{I} \sum_{j}^{J} \sum_{l=1}^{n_{ij}} \left(Y_{ijl} - \bar{Y}_{ij\cdot}\right)^2$$

```
> ## Calculate SS_E
> S_E = aggregate(Obs ~ cycle_times + operator, data = Obs_w_names,
                  FUN = function(vec) sum((vec-mean(vec))^2))
> head(S_E, 6)
  cycle_times operator    Obs
1          40        1  2.000
2          50        1  0.667
3          60        1  8.667
4          40        2  2.000
5          50        2 14.000
6          60        2  0.667

> SS_E = sum(S_E$Obs)
> SS_E
[1] 36.66667
```

Two treatment factors

rep(a, times = 3) rep(a, each = 3)
a = c(1,2,3)
times ⇒ 1, 2, 3, 1, 2, 3, 1, 2, 3
each ⇒ 1, 1, 1, 2, 2, 2, 3, 3, 3

Example 4 *cont'd*. A fabric plant is studying the effect of several factors on the dyeing of cotton-synthetic cloth. At 300°C, three operators and three cycle times were selected. The finished cloth was compared to a standard, and a numerical score was assigned.

Operators

Cycle times	1	2	3
40 mins	23	27	31
	24	28	32
	25	26	29
50 mins	36	34	33
	35	38	34
	36	39	35
60 mins	28	35	26
	24	35	27
	27	34	25

```
> head(Obs_w_names, 6)
  cycle_times operator Obs
1          40        1  23
2          40        1  24
3          40        1  25
4          50        1  36
5          50        1  35
6          50        1  36
```

$$SS_{AB} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots}\right)^2$$

```
> ## Calculate SS_AB
> cell_mean = aggregate(Obs ~ cycle_times + operator, data = Obs_w_names,
                        FUN = mean)
> cell_mean_vec = rep(cell_mean$Obs, each = p)

> A_means =  aggregate(Obs ~ cycle_times, data = Obs_w_names,
                       FUN = mean)
> A_means_vec = rep(rep(A_means$Obs, times = J), each = p)

> B_means =  aggregate(Obs ~ operator, data = Obs_w_names,
                       FUN = mean)
> B_means_vec = rep(rep(B_means$Obs, each = I), each = p)

> Tmp_data = cbind(cell_mean_vec, A_means_vec, B_means_vec, all_mean)
> S_AB = apply(Tmp_data, 1, function(vec) (vec[1]-vec[2]-vec[3]+vec[4])^2)
> SS_AB = sum(S_AB)
> SS_AB
[1] 143.037
```

Two treatment factors

Example 4 *cont'd*. A fabric plant is studying the effect of several factors on the dyeing of cotton-synthetic cloth. At 300°C, three operators and three cycle times were selected. The finished cloth was compared to a standard, and a numerical score was assigned.

$H_0 : \alpha_1 = \cdots = \alpha_I = 0$

Reject H_0

```
> ## Test main effect of A
> F_stat = {SS_A/(I-1)}/{SS_E/(n-I*J)}
> F_stat
[1] 85.10909

> # Critical value
> qf(0.05, I-1, n-I*J, lower.tail = FALSE)
[1] 3.554557

> # p-value
> pf(F_stat, I-1, n-I*J, lower.tail = FALSE)
[1] 6.691113e-10
```

$H_0 : \beta_1 = \cdots = \beta_J = 0$

Reject H_0

```
> ## Test main effect of B
> F_stat = {SS_B/(J-1)}/{SS_E/(n-I*J)}
> F_stat
[1] 20.14545

> # Critical value
> qf(0.05, J-1, n-I*J, lower.tail = FALSE)
[1] 3.554557

> # p-value
> pf(F_stat, J-1, n-I*J, lower.tail = FALSE)
[1] 2.552966e-05
```

$H_0 :$ all δ_{ij}'s are zero

Reject H_0

```
> ## Test interaction A*B
> F_stat = {SS_AB/((I-1)*(J-1))}/{SS_E/(n-I*J)}
> F_stat
[1] 17.55455

> # Critical value
> qf(0.05, (I-1)*(J-1), n-I*J, lower.tail = FALSE)
[1] 2.927744

> # p-value
> pf(F_stat, (I-1)*(J-1), n-I*J, lower.tail = FALSE)
[1] 5.003958e-06
```

Two treatment factors

Example 4 *cont'd*. A fabric plant is studying the effect of several factors on the dyeing of cotton-synthetic cloth. At 300°C, three operators and three cycle times were selected. The finished cloth was compared to a standard, and a numerical score was assigned.

```
> fit <- lm(Obs ~ cycle_times*operator, data = Obs_w_names)
> anova(fit)
Analysis of Variance Table

Response: Obs
                    Df Sum Sq Mean Sq F value    Pr(>F)
cycle_times          2 346.74 173.370  85.109 6.691e-10 ***
operator             2  82.07  41.037  20.145 2.553e-05 ***
cycle_times:operator 4 143.04  35.759  17.555 5.004e-06 ***
Residuals           18  36.67   2.037
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obs ~ cycle_times + operator + cycle_times : operator

Two treatment factors

Model 1: $y_{ijl} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijl}$

Model 2: $y_{ijl} = \mu + \alpha_i + \beta_j + \epsilon_{ijl}$

model 2: $y_{ijl} = \mu + \alpha_i$

H_0: Model 1 is worse than model 2.

Example 4 *cont'd*. A fabric plant is studying the effect of several factors on the dyeing of cotton-synthetic cloth. At 300°C, three operators and three cycle times were selected.

```
> fit <- lm(Obs ~ cycle_times*operator, data = Obs_w_names)
> anova(fit)
Analysis of Variance Table

Response: Obs
                   Df Sum Sq Mean Sq F value    Pr(>F)
cycle_times         2 346.74 173.370  85.109 6.691e-10 ***
operator            2  82.07  41.037  20.145 2.553e-05 ***
cycle_times:operator  4 143.04  35.759  17.555 5.004e-06 ***
Residuals          18  36.67   2.037
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model comparisons

H_0: $\beta_j = 0$, $\delta_{ij} = 0$

```
> # F test as a model comprarison
> fit_0 <- lm(Obs ~ cycle_times + operator, data = Obs_w_names)
> anova(fit_0, fit, test='F')
Analysis of Variance Table

Model 1: Obs ~ cycle_times + operator
Model 2: Obs ~ cycle_times * operator
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     22 179.704
2     18  36.667  4    143.04 17.555 5.004e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

∈ Model 1

∈ Model 2

```
> fit_0 <- lm(Obs ~ cycle_times, data = Obs_w_names)
> anova(fit_0, fit, test='F')
Analysis of Variance Table

Model 1: Obs ~ cycle_times
Model 2: Obs ~ cycle_times * operator
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     24 261.778
2     18  36.667  6    225.11 18.418 8.719e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One-way MANOVA

Some basics

07/29/2021

Multivariate analysis of variance (MANOVA)

The one-way layout involves **one factor** in the experimental design:

Treatments				
1	2	3	...	k
$\mathbf{y}_{11}$	$\mathbf{y}_{21}$	$\mathbf{y}_{31}$	$\cdots$	$\mathbf{y}_{k1}$
$\mathbf{y}_{12}$	$\mathbf{y}_{22}$	$\mathbf{y}_{32}$	$\cdots$	$\mathbf{y}_{k2}$
$\vdots$	$\vdots$	$\vdots$	...	$\mathbf{y}_{k3}$
		$\mathbf{y}_{3n_3}$		$\vdots$
$\mathbf{y}_{1n_1}$				
	$\mathbf{y}_{2n_2}$			$\mathbf{y}_{kn_k}$

$$\mathbf{y} = (y^1, \cdots, y^p)$$

Model assumption $(j = 1, \cdots, n_i,\ i = 1, \cdots, k)$:

$$\boldsymbol{Y}_{ij} = \underbrace{\boldsymbol{\mu}}_{\text{common mean level}} + \underbrace{\boldsymbol{\alpha}_i}_{\text{unique effect due to treatment } i} + \underbrace{\boldsymbol{\epsilon}_{ij}}_{\overset{\text{iid}}{\sim} N(\mathbf{0}, \Sigma)}.$$

Gaussian

$$H_0 : \boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_k = 0 \quad \text{versus} \quad H_1 : \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j \text{ for some } i \neq j.$$

One-way MANOVA: LRT

Lemma G. The LRT rejection region is solely based on the following identity:

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{\cdot\cdot}\right)\left(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{\cdot\cdot}\right)'}_{\text{Total sum of squares}}=\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{i\cdot}\right)\left(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{i\cdot}\right)'}_{\text{SS within groups}}+\underbrace{\sum_{i=1}^{k}n_i\left(\bar{\boldsymbol{Y}}_{i\cdot}-\bar{\boldsymbol{Y}}_{\cdot\cdot}\right)\left(\bar{\boldsymbol{Y}}_{i\cdot}-\bar{\boldsymbol{Y}}_{\cdot\cdot}\right)'}_{\text{SS between groups}}$$

$$\mathbf{T} = \mathbf{E} + \mathbf{B}$$

Treatments				
1	2	3	...	k
$\mathbf{y}_{11}$	$\mathbf{y}_{21}$	$\mathbf{y}_{31}$	$\cdots$	$\mathbf{y}_{k1}$
$\mathbf{y}_{12}$	$\mathbf{y}_{22}$	$\mathbf{y}_{32}$	$\cdots$	$\mathbf{y}_{k2}$
$\vdots$	$\vdots$	$\vdots$	...	$\mathbf{y}_{k3}$
		$\mathbf{y}_{3n_3}$		$\vdots$
$\mathbf{y}_{1n_1}$				
	$\mathbf{y}_{2n_2}$			$\mathbf{y}_{kn_k}$

Proof.

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{i\cdot}+\bar{\boldsymbol{Y}}_{i\cdot}-\bar{\boldsymbol{Y}}_{\cdot\cdot}\right)\left(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{i\cdot}+\bar{\boldsymbol{Y}}_{i\cdot}-\bar{\boldsymbol{Y}}_{\cdot\cdot}\right)'$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{i\cdot}\right)\left(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{i\cdot}\right)'$$

$$+\sum_{i=1}^{k}n_i\left(\bar{\boldsymbol{Y}}_{i\cdot}-\bar{\boldsymbol{Y}}_{\cdot\cdot}\right)\left(\bar{\boldsymbol{Y}}_{i\cdot}-\bar{\boldsymbol{Y}}_{\cdot\cdot}\right)'$$

$$+2\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}\left(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{i\cdot}\right)\right]\left(\bar{\boldsymbol{Y}}_{i\cdot}-\bar{\boldsymbol{Y}}_{\cdot\cdot}\right)' \quad \left[\sum_{j=1}^{n_i}\left(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{i\cdot}\right)\right]=0$$

One-way MANOVA: LRT

Proposition G. There are many different test statistics:

1) Wilk's Lambda: $\mathbf{\Lambda}^* = \dfrac{|\mathbf{E}|}{|\mathbf{B}+\mathbf{E}|}$.

2) Hotelling-Lawley trace: $T_0^2 = \text{trace}(\mathbf{B}\mathbf{E}^{-1})$.

3) Pillai trace: $V = \text{trace}(\mathbf{B}(\mathbf{B}+\mathbf{E})^{-1})$.

4) Roy's maximum root: the largest eigenvalue of $\mathbf{B}\mathbf{E}^{-1}$.

MANOVA

Source	d.f.	SSP
Treatments	*k - 1*	**B**
Error	*n - k*	**E**
Total	*n - 1*	**T**

In the univariate case,

$$\lambda(\mathbf{Y}_n) = \left(\frac{SS_W}{SS_{\text{Tot}}}\right)^{n/2},$$

$$R = \left\{\frac{SS_B}{SS_W} \geq c''\right\}.$$

The distribution of these statistics under H_0 is not straightforward and can only be approximated except when p is small.

One-way MANOVA: LRT

Example 5. A researcher randomly assigns 33 subjects to one of 3 groups that receive: 1) technical dietary information interactively from an online website; 2) the same information from a nurse practitioner; 3) the information from a video tape made by the same nurse practitioner. The researcher looks at three different ratings of the presentation, difficulty, usefulness and importance, to determine if there is a difference in the modes of presentation.

	Group	Useful	Difficulty	Importance
1	1	19.6	5.15	9.5
2	1	15.4	5.75	9.1
3	1	22.3	4.35	3.3
⋮				
12	2	17.1	9.00	7.5
13	2	15.7	5.30	8.5
14	2	14.9	9.85	6.0
⋮				

```
# MANOVA test
res.man <- manova(cbind(Useful, Difficulty, Importance) ~ factor(Group),
                  data = dataset)
summary(res.man)
```

```
                Df  Pillai approx F num Df den Df  Pr(>F)
factor(Group)    2 0.47667   3.0248      6     58 0.01215 *
Residuals       30
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

One-way MANOVA: LRT

```
> summary(res.man, 'Wilks')
             Df   Wilks approx F num Df den Df   Pr(>F)
factor(Group)  2 0.52579   3.5382      6     56 0.004859 **
Residuals     30
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(res.man, 'Hotelling-Lawley')
             Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
factor(Group)  2          0.89723   4.0375      6     54 0.002058 **
Residuals     30
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(res.man, 'Roy')
             Df     Roy approx F num Df den Df    Pr(>F)
factor(Group)  2 0.89199   8.6225      3     29 0.0003023 ***
Residuals     30
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
# MANOVA test
res.man <- manova(cbind(Useful, Difficulty, Importance) ~ factor(Group),
                  data = dataset)
summary(res.man)
```

```
                Df  Pillai approx F num Df den Df  Pr(>F)
factor(Group)  2 0.47667   3.0248      6     58 0.01215 *
Residuals     30
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Next Tuesday ...

Simple linear regression:

- Model assumptions;
- Geometric interpretations;
- Maximum likelihood estimation.