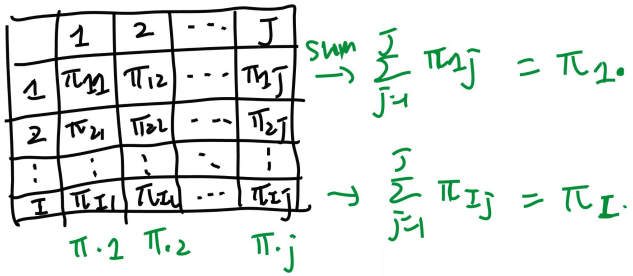
Analysis of Variance (ANOVA)

Chapter 12 of Rice

07/27/2021

Berkeley
UNIVERSITY OF CALIFORNIA

In the previous lecture,

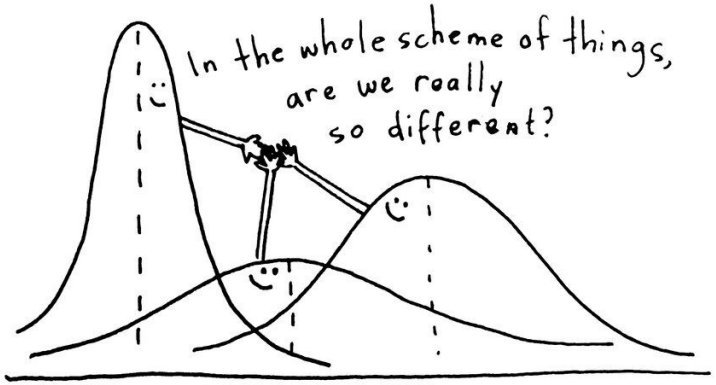


- χ^2 test of independence:
  - For a 2x2 table, we can use Fisher's exact test;
  - For a I x J table, we assume under the null hypothesis that
    $$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}, \ \ i = 1, \ldots, I, \ \ \ j = 1, \ldots, J.$$
  - The difference between the numbers of free parameters:
    $$\nu = (IJ - 1) - (I + J - 2) = (I - 1)(J - 1).$$
- χ^2 test of homogeneity:
  - We assume under the null hypothesis that
    $$H_0 : \pi_{i1} = \pi_{i2} = \cdots = \pi_{iJ}, \ \ i = 1, \ldots, I.$$
  - The difference between the numbers of free parameters:
    $$\nu = J(I - 1) - (I - 1) = (I - 1)(J - 1).$$
- In both case
  - Under H_0,
    $$-2 \log \lambda(\mathbf{X}_n) \xrightarrow{d} \chi^2_{(I-1)(J-1)} \ \text{ as } n \to \infty.$$

Analysis of variance

Example 1. To compare the effects of three toxins and a control on the liver of a certain species of trout, researchers recorded the amounts of deterioration (in standard units) of the liver of each sacrificed fish.

Toxin 1	Toxin 2	Toxin 3	Control
28	33	18	11
23	36	21	14
14	34	20	11
27	29	22	16
	31	24	
	34		



ANOVA: Comparing the means of several populations via analyzing variances.

One-way vs Two-way ANOVA

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

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

Main interest: *difference in means*

The two-way layout involves **two factors** in the experimental design:

	1	2	$\cdots$	J
1	$y_{111}, y_{112}, \cdots, y_{11n_{11}}$	$y_{121}, y_{122}, \cdots, y_{12n_{12}}$	$\cdots$	$y_{1J1}, y_{1J2}, \cdots, y_{1Jn_{1J}}$
2	$y_{211}, y_{212}, \cdots, y_{21n_{21}}$	$y_{221}, y_{222}, \cdots, y_{22n_{22}}$	$\cdots$	$y_{2J1}, y_{2J2}, \cdots, y_{2Jn_{2J}}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I				

⇐ 2nd Factor

⇑ 1st Factor

$y_{ijk},\ k = 1, \cdots, n_{ij}.$

Main interest: *difference in means + interrelationship of the two factors*

Multivariate analysis of variance (MANOVA)

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

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

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

The two-way layout involves **two factors** in the experimental design:

	1	2	$\cdots$	J
1	$\mathbf{y}_{111}, \mathbf{y}_{112}, \cdots, \mathbf{y}_{11n_{11}}$	$\mathbf{y}_{121}, \mathbf{y}_{122}, \cdots, \mathbf{y}_{12n_{12}}$	$\cdots$	$\mathbf{y}_{1J1}, \mathbf{y}_{1J2}, \cdots, \mathbf{y}_{1Jn_{1J}}$
2	$\mathbf{y}_{211}, \mathbf{y}_{212}, \cdots, \mathbf{y}_{21n_{21}}$	$\mathbf{y}_{221}, \mathbf{y}_{222}, \cdots, \mathbf{y}_{22n_{22}}$	$\cdots$	$\mathbf{y}_{2J1}, \mathbf{y}_{2J2}, \cdots, \mathbf{y}_{2Jn_{2J}}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I				

⇐ 2nd Factor

⇑ 1st Factor

$\mathbf{y}_{ijk},\ k = 1, \cdots, n_{ij}.$

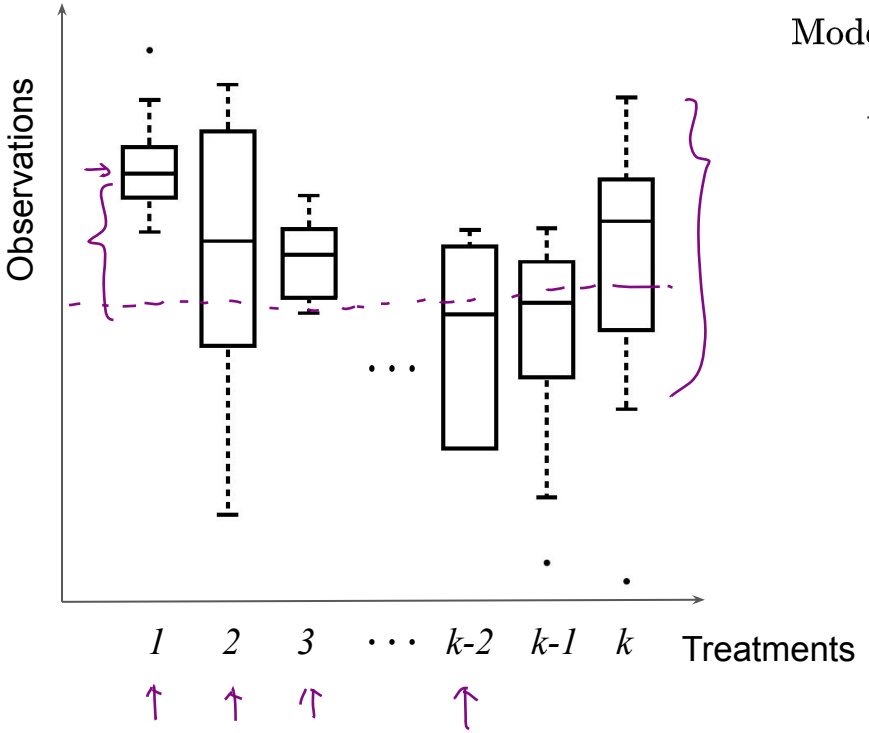
One-way ANOVA (univariate)

12.2 of Rice

07/27/2021

One-way ANOVA: model assumption

H_0 : All population means are the same.



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

$$Y_{ij} = \underbrace{\mu}_{\text{common mean level}} + \underbrace{\alpha_i}_{\text{unique effect due to treatment } i} + \underbrace{\epsilon_{ij}}_{\overset{\text{iid}}{\sim} N(0,\sigma^2)} .$$

Note $\sum_{i=1}^{k} \alpha_i = 0.$

$$H_0 : \alpha_1 = \cdots = \alpha_k = 0 \quad \text{versus} \quad H_1 : \alpha_i \neq \alpha_j \text{ for some } i \neq j$$

$Y_{ij} = \mu + \epsilon_{ij}$

$\overset{iid}{\sim} N(\mu, \sigma^2)$

One-way ANOVA: LRT

Example 2. Assume $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, where $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. We can derive the LRT for the hypotheses:

$$H_0 : \alpha_1 = \cdots = \alpha_k \quad \text{versus} \quad H_1 : \alpha_i = \alpha_j \text{ for some } i \neq j.$$

Solution. $\Theta_0 = \{\alpha_1 = \cdots = \alpha_k = 0, \ \mu \in \mathbb{R}, \ \sigma^2 > 0\}$, $\Theta = \{\sum_{i=1}^k \alpha_i = 0, \ \mu \in \mathbb{R}, \ \sigma^2 > 0\}$

$\dim \Theta_0 = 1+1$

$\dim \Theta = k-1+1+1 = k+1$

$\to \nu = k-1$

$n = \sum_{i=1}^k n_i$

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

$Y_{ij} \overset{\text{iid}}{\sim} N(\mu + \alpha_i, \sigma^2)$

$L(\mu, \alpha_1, \cdots \alpha_k \mid Y_{ij}) = \prod_{i=1}^k \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_{ij} - \mu - \alpha_i)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^k \sum_{j=1}^{n_i} \frac{(Y_{ij} - \mu - \alpha_i)^2}{2\sigma^2}}$

$\sup_{\Theta_0} L(\mu, \alpha_1, \cdots, \alpha_k, \sigma^2 \mid Y_{ij})$

$= \sup_{\substack{\mu \in \mathbb{R} \\ \sigma^2 > 0}} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^k \sum_{j=1}^{n_i} \frac{(Y_{ij} - \mu)^2}{2\sigma^2}}$

$\hat{\mu}_0 = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}}{n} = \bar{Y}_{..}$, $\hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$

$\Rightarrow \sup_{\Theta_0} L(\mu, \alpha_1, \cdots \alpha_k \mid Y_{ij}) = \left(\frac{1}{\sqrt{2\pi\hat{\sigma}_0^2}}\right)^n e^{-\frac{n}{2}}$ $\Leftarrow$

$\prod_{j=1}^{n_1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_{ij} - \mu - \alpha_1)^2}{2\sigma^2}}$

One-way ANOVA: LRT

Solution *cont'd*.

$$\sup_{H} L(\mu, \alpha_1, \cdots \alpha_k, b^2 \mid Y_{ij}) = \sup_{H} \left(\frac{1}{\sqrt{2\pi b^2}}\right)^n e^{-\frac{1}{2b^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\mu-\alpha_i)^2} \Leftarrow$$

$$l(\underbrace{\mu, \alpha_1, \cdots \alpha_k}_{k+1}, \underbrace{b^2}_{1} \mid Y_{ij}) = -\frac{n}{2}\log(2\pi b^2) - \frac{1}{2b^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\mu-\alpha_i)^2 \Leftarrow \quad -\alpha_k = \alpha_1 + \cdots + \alpha_{k-1}$$

$$= -\frac{n}{2}\log(2\pi b^2) - \frac{1}{2b^2}\sum_{i=1}^{k-1}\sum_{j=1}^{n_i}(Y_{ij}-\mu-\alpha_i)^2 - \frac{1}{2b^2}\sum_{j=1}^{n_k}(Y_{kj}-\mu+\alpha_1+\cdots+\alpha_{k-1})^2$$

$$\begin{cases}
\frac{\partial l}{\partial \mu} = \frac{1}{b^2}\sum_{i=1}^{k-1}\sum_{j=1}^{n_i}(Y_{ij}-\mu-\alpha_i) + \frac{1}{b^2}\sum_{j=1}^{n_k}(Y_{kj}-\mu+\underbrace{\alpha_1+\cdots+\alpha_{k-1}}_{-\alpha_k}) = 0 \Leftarrow \\
\frac{\partial l}{\partial \alpha_1} = \frac{1}{b^2}\sum_{j=1}^{n_1}(Y_{1j}-\mu-\alpha_1) - \frac{1}{b^2}\sum_{j=1}^{n_k}(Y_{kj}-\mu+\alpha_1+\cdots+\alpha_{k-1}) = 0 \Leftarrow \\
\vdots \\
\frac{\partial l}{\partial \alpha_{k-1}} = \frac{1}{b^2}\sum_{j=1}^{n_{k-1}}(Y_{k-1,j}-\mu-\alpha_{k-1}) - \frac{1}{b^2}\sum_{j=1}^{n_k}(Y_{kj}-\mu+\alpha_1+\cdots+\alpha_{k-1}) = 0 \\
\frac{\partial l}{\partial b} = -\frac{n}{b} + \frac{1}{b^3}\sum_{i=1}^{k}(Y_{ij}-\mu-\alpha_i)^2 = 0 \Leftarrow
\end{cases}$$

From the first equation, we get $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\mu-\alpha_i) = 0 \Leftarrow$

Summing over the $k-1$ equations in the middle, we get $\frac{1}{b^2}\sum_{i=1}^{k-1}\sum_{j=1}^{n_i}(Y_{ij}-\mu-\alpha_i) - \frac{1}{b^2}\sum_{j=1}^{n_k}(Y_{kj}-\mu-\alpha_k) = 0$

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\mu-\alpha_i) - \sum_{j=1}^{n_k}(Y_{kj}-\mu-\alpha_k) = 0$

One-way ANOVA: LRT

Solution *cont'd*.

population mean for k-th treatment

which means $\sum_{j=1}^{n_k}(Y_{kj}-\mu-\alpha_k)=0 \Rightarrow \mu+\alpha_k=\frac{\sum_{j=1}^{n_k}Y_{kj}}{n_k}=\bar{Y}_{k\cdot}$

From the 2nd equation, $\frac{1}{\sigma^2}\sum_{j=1}^{n_1}(Y_{1j}-\mu-\alpha_1)-\frac{1}{\sigma^2}\sum_{j=1}^{n_k}(Y_{kj}-\mu-\alpha_k)=0$

$\Rightarrow \widehat{\mu+\alpha_1}=\frac{\sum_{j=1}^{n_1}Y_{1j}}{n_1}=\bar{Y}_{1\cdot}$

Similarly, $\widehat{\mu+\alpha_i}=\frac{\sum_{j=1}^{n_i}Y_{ij}}{n_i}=\bar{Y}_{i\cdot}$

$\hat{\sigma}_n^2=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\mu-\alpha_i)^2=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\cdot})^2$

Thus, $\sup_{\Theta} L(\mu,\alpha_1,\cdots\alpha_k,\sigma^2|Y_{ij})=\left(\frac{1}{2\pi\hat{\sigma}_n^2}\right)^{\frac{n}{2}}e^{-\frac{n}{2}}$.

$\lambda(Y_n)=\frac{\sup_{\Theta_0}L}{\sup_{\Theta}L}=\left(\frac{\hat{\sigma}_n^2}{\hat{\sigma}_0^2}\right)^{\frac{n}{2}}$

$-2\log\lambda(Y_n)\xrightarrow{d}\chi_\nu^2=\chi_{k-1}^2$ under H_0.

One-way ANOVA: LRT

$R = \{\lambda(Y_n) \le c\} = \left\{\left(\frac{\hat\sigma_n^2}{\hat\sigma_0^2}\right)^{\frac{n}{2}} \le c\right\} = \left\{\frac{SS_W}{SS_{tot}} \le c^{\frac{2}{n}}\right\}$ $\hat\sigma_n^2$ $SS_W + SS_B$

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

$$\hat\sigma_0^2 = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar Y_{\cdot\cdot}\right)^2}_{\text{Total sum of squares}} = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar Y_{i\cdot}\right)^2}_{\text{SS within groups}} + \underbrace{\sum_{i=1}^{k} n_i\left(\bar Y_{i\cdot}-\bar Y_{\cdot\cdot}\right)^2}_{\text{SS between groups}}.$$

SS_{tot} SS_W SS_B

$= \left\{\frac{SS_B}{SS_W} \ge c'\right\}$

Proof. $\hat\sigma_0^2 - \hat\sigma_n^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar Y_{\cdot\cdot})^2 - \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar Y_{i\cdot})^2$

$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}^2 + \bar Y_{\cdot\cdot}^2 - 2Y_{ij}\bar Y_{\cdot\cdot} - Y_{ij}^2 - \bar Y_{i\cdot}^2 + 2Y_{ij}\bar Y_{i\cdot}\right)$

$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar Y_{\cdot\cdot}^2 - 2Y_{ij}\bar Y_{\cdot\cdot} - \bar Y_{i\cdot}^2 + 2Y_{ij}\bar Y_{i\cdot}\right)$

$= \sum_{i=1}^{k}\left(n_i\bar Y_{\cdot\cdot}^2 - 2\bar Y_{\cdot\cdot}\sum_{j=1}^{n_i}Y_{ij} - n_i\bar Y_{i\cdot}^2 + 2\bar Y_{i\cdot}\sum_{j=1}^{n_i}Y_{ij}\right)$

$\sum_{j=1}^{n_i}Y_{ij} = n_i\bar Y_{i\cdot}$

$= \sum_{i=1}^{k} n_i\left[\bar Y_{\cdot\cdot}^2 - 2\bar Y_{\cdot\cdot}\bar Y_{i\cdot} - \bar Y_{i\cdot}^2 + 2\bar Y_{i\cdot}^2\right]$ $\bar Y_{i\cdot}^2$

$= \sum_{i=1}^{k} n_i\left(\bar Y_{\cdot\cdot} - \bar Y_{i\cdot}\right)^2.$

Treatments				
1	2	3	...	k
y_{11}	y_{21}	y_{31}	...	y_{k1}
y_{12}	y_{22}	y_{32}	...	y_{k2}
⋮	⋮	⋮	...	y_{k3}
		y_{3n_3}		⋮
y_{1n_1}				
	y_{2n_2}			y_{kn_k}

$\bar Y_{1\cdot}$ $\bar Y_{2\cdot}$ $\bar Y_{k\cdot}$ $\sum_{j=1}^{n_1}(Y_{1j}-\bar Y_{1\cdot})^2$

One-way ANOVA: LRT

$\chi^2_n + \chi^2_m = \chi^2_{n+m}$

Proposition A. Under the one-way ANOVA model assumption,

$$\frac{SS_W}{\sigma^2} \sim \chi^2_{n-k}, \text{ and it's independent of each } \bar{Y}_{i\cdot},\ i = 1, \ldots, k.$$

Furthermore, under H_0, $\frac{SS_B}{\sigma^2} \sim \chi^2_{k-1}$, and it's independent of SS_W.

Proof*. $\frac{1}{\sigma^2}\sum_{j=1}^{n_1}(Y_{1j} - \bar{Y}_{1\cdot})^2 \sim \chi^2_{n_1-1}$ $\perp\!\!\!\perp$ $\bar{Y}_{1\cdot}$

$\vdots$

$\frac{1}{\sigma^2}\sum_{j=1}^{n_k}(Y_{kj} - \bar{Y}_{k\cdot})^2 \sim \chi^2_{n_k-1}$ $\perp\!\!\!\perp$ $\bar{Y}_{k\cdot}$

$Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2}$ $\overset{iid}{\sim} N(\mu+\alpha_1, \sigma^2)$

Summing over i, we have

$$\frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_k}(Y_{ij} - \bar{Y}_{i\cdot})^2 \sim \chi^2_{\sum_{i=1}^{k}(n_i-1)} = \chi^2_{n-k}. \quad \perp\!\!\!\perp \bar{Y}_{1\cdot}, \ldots \bar{Y}_{k\cdot}.$$

$\frac{1}{\sigma^2}SS_W$ $\qquad SS_B$ $\perp\!\!\!\perp$

One-way ANOVA: LRT

Proof* *cont'd.*

$SS_B = \sum_{i=1}^{k} n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \Rightarrow \frac{1}{\sigma^2}(Z_2^2 + \cdots + Z_k^2) = \frac{1}{\sigma^2}\left[\sum_{i=1}^{k} n_i (\bar{Y}_{i\cdot} - \mu)^2 - n(\bar{Y}_{\cdot\cdot} - \mu)^2\right] \sim \chi^2_{k-1}$

We know that $\bar{Y}_{i\cdot} \sim N\left(\mu + \alpha_i, \frac{\sigma^2}{n_i}\right)$, which is $\sqrt{n_i}(\bar{Y}_{i\cdot} - \mu - \alpha_i) \sim N(0, \sigma^2)$.

Under H_0, $\begin{pmatrix} \sqrt{n_1}(\bar{Y}_{1\cdot} - \mu) \\ \vdots \\ \sqrt{n_k}(\bar{Y}_{k\cdot} - \mu) \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \underbrace{\begin{pmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{pmatrix}}_{=\sigma^2 I}\right]$

$\Rightarrow \frac{SS_B}{\sigma^2} \sim \chi^2_{k-1}$ ⫫ $\bar{Y}_{\cdot\cdot}$

We want to construct an orthogonal matrix such that

$\begin{pmatrix} Z_1 \\ \vdots \\ Z_k \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\sqrt{n_1}}{\sqrt{n}} & \cdots & \frac{\sqrt{n_k}}{\sqrt{n}} \\ - & - & - \\ - & - & - \\ - & - & - \end{pmatrix}}_{A} \begin{pmatrix} \sqrt{n_1}(\bar{Y}_{1\cdot} - \mu) \\ \vdots \\ \sqrt{n_k}(\bar{Y}_{k\cdot} - \mu) \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma^2 \underbrace{AIA^T}_{=AA^T=I}\right] = N\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma^2 I\right)$

$\frac{1}{\sigma^2}(Z_2^2 + \cdots + Z_k^2) \sim \chi^2_{k-1}$ ⫫ $Z_1 = \frac{\sqrt{n_1}}{\sqrt{n}} \times \sqrt{n_1}(\bar{Y}_{1\cdot} - \mu) + \cdots + \frac{\sqrt{n_k}}{\sqrt{n}} \times \sqrt{n_k}(\bar{Y}_{k\cdot} - \mu)$

$= \frac{1}{\sqrt{n}} \sum_{i=1}^{k} n_i (\bar{Y}_{i\cdot} - \mu) = \frac{\sqrt{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu)}{n} = \sqrt{n}(\bar{Y}_{\cdot\cdot} - \mu)$

One-way ANOVA: sampling distribution under H_0

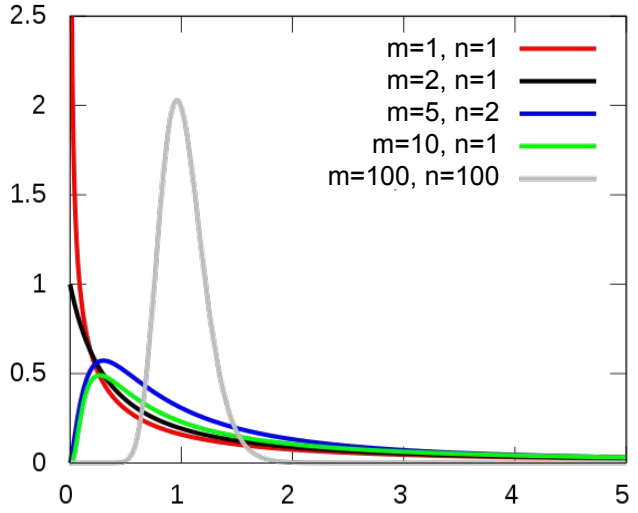
Definition. Let $U \sim \chi^2_m$, $V \sim \chi^2_n$ and $U \perp\!\!\!\perp V$. Then

$$\frac{U/m}{V/n} \sim F_{m,n},$$

which is called a *F* distribution with m and n degrees of freedom.

$$f(w) = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1+\frac{m}{n}w\right)^{-(m+n)/2}, \qquad w \geq 0$$

If $X \sim t_n$, then $X^2 \sim F_{1,n}$.



One-way ANOVA: sampling distribution under H_0

Theorem A. Under the one-way ANOVA model assumption, the LRT rejection region is equivalent to

$$R = \left\{ \frac{SS_B/(k-1)}{SS_W/(n-k)} \geq c'' \right\}.$$

Under H_0, $\frac{SS_B/(k-1)}{SS_W/(n-k)} \sim F_{k-1,\, n-k}$.

Proof. See previous page.

To ensure the significance level $= \alpha$,

$$c'' = qf\left(\alpha,\ df1 = k-1,\ df2 = n-k,\ lower.tail = FALSE\right)$$

$$p\text{-value} = pf\left(\frac{SS_B/(k-1)}{SS_W/(n-k)},\ df1 = k-1,\ df2 = n-k,\ lower.tail = FALSE\right).$$

One-way ANOVA: summary table

F statistic $= \dfrac{SS_B/(k-1)}{SS_W/(n-k)}$

Source	df	Sum Sq	Mean Sq	*F* value	*p*-value
Between group (*or treatment*)	$k-1$	$SS_B = \sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$MS_B = SS_B/(k-1)$	MS_B / MS_W	`pf(MS_B/MS_W, k-1, n-k, lower.tail = FALSE)`
Within group (*or residuals*)	$n-k$	$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$	$MS_W = SS_W/(n-k)$		
Total	$n-1$	$SS_{Total} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$			

One-way ANOVA: implementation

tmp = list(c(1,2), c(2,3,4))
tmp
[[1]] 1, 2
[[2]] 2, 3, 4
unlist(tmp) 1, 2, 2, 3, 4

Example 1. To compare the effects of three toxins and a control on the liver of a certain species of trout, researchers recorded the amounts of deterioration (in standard units) of the liver of each sacrificed fish.

```
group1 <- c(28, 23, 14, 27)
group2 <- c(33, 36, 34, 29, 31, 34)
group3 <- c(18, 21, 20, 22, 24)
group4 <- c(11, 14, 11, 16)
#Combine into a list
Observations <- list(group1, group2, group3, group4)
k <- length(Observations); n <- length(unlist(Observations)) #bookkeeping
```

Read in data

```
# Squares within groups
S_W <- lapply(Observations, function(vec) sum((vec-mean(vec))^2))
print(unlist(S_W))
[1] 122.000  30.833  20.000  18.000

# Squares between groups
tot_mean <- mean(unlist(Observations))
S_B <- lapply(Observations, function(vec) length(vec)*(mean(vec)-tot_mean)^2)
print(unlist(S_B))
[1] 0.898 525.618  30.596 438.792
```

```
# F statistic
SS_B <- sum(unlist(S_B))
SS_B
[1] 995.9035

SS_W <- sum(unlist(S_W))
SS_W
[1] 190.8333

{SS_B/(k-1)}/{SS_W/(n-k)}
[1] 26.09354

# Critical value
qf(0.95, k-1, n-k, lower.tail = FALSE)
[1] 0.1149046
```

```
# p-value
pf(26.09354, k-1, n-k, lower.tail = FALSE)
[1] 3.347973e-06
```

One-way ANOVA: implementation

Example 1. To compare the effects of three toxins and a control on the liver of a certain species of trout, researchers recorded the amounts of deterioration (in standard units) of the liver of each sacrificed fish.

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

linear population mean

linear model

```
> fit <- lm(values ~ ind, data = input)
> anova(fit)
Analysis of Variance Table

Response: values
          Df Sum Sq Mean Sq F value    Pr(>F)
ind        3 995.90  331.97  26.093 3.348e-06 ***
Residuals 15 190.83   12.72
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> group_ind <- c('toxin1', 'toxin2', 'toxin3', 'control')
> input <- stack(setNames(Observations, group_ind))
> head(input, 6)
  values    ind
1     28 toxin1
2     23 toxin1
3     14 toxin1
4     27 toxin1
5     33 toxin2
6     36 toxin2
```

One-way ANOVA (univariate)

Simultaneous estimates of contrasts

07/27/2021

Silly null hypothesis

Example 1. To compare the effects of three toxins and a control on the liver of a certain species of trout, researchers recorded the amounts of deterioration (in standard units) of the liver of each sacrificed fish.

$$H_0 : \alpha_1 = \cdots = \alpha_k = 0.$$

Toxin 1	Toxin 2	Toxin 3	Control
28	33	18	11
23	36	21	14
14	34	20	11
27	29	22	16
	31	24	
	34		

The ANOVA test gives no information about **how** they differ, in particular about **which pairs** are significantly different.

```
# p-value
pf(26.09354, k-1, n-k, lower.tail = FALSE)
[1] 3.347973e-06
```

Simultaneous Bonferroni CIs

Theorem B. Under the one-way ANOVA model assumptions,

$$\frac{\sqrt{n_i}\left(\bar{Y}_{i\cdot} - \mu - \alpha_i\right)}{S_i} \sim t_{n_i - 1}, \text{ for any } i = 1, \ldots, k.$$

If one establishes m confidence intervals, and wishes to have an overall confidence level of $1 - \alpha$, each individual confidence interval can be adjusted to the level of $1 - \frac{\alpha}{m}$.

Simultaneous $(1 - \alpha) \times 100\%$ CIs for population means :

$$\bar{Y}_{i\cdot} \pm t_{n_i - 1}\left(\alpha'/2\right)\frac{S_i}{\sqrt{n_i}} \quad \text{for any } i = 1, \ldots, k,$$

where $\alpha' = \cancel{2\alpha/[k(k-1)]}$.

α/k

```r
get_bound <- function(vec, k, alpha = 0.05){
  center = mean(vec)
  n = length(vec)
  halfwidth = qt(1 - (alpha/2)/k, df = n - 1)*sd(vec)/sqrt(n)

  return(c(center - halfwidth, center + halfwidth))
}
```

Simultaneous inference statements

Corollary B. We can utilize the duality between the CIs and the HTs to perform testings on

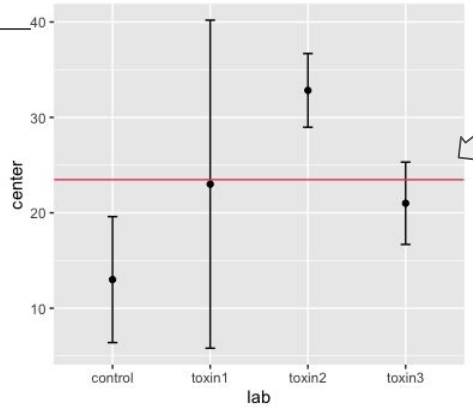
$$H_0^i : \alpha_i = 0 \quad \text{versus} \quad H_1^i : \alpha_i \neq 0.$$

The conclusions we draw about each test hold **simultaneously**.

Toxin 1	Toxin 2	Toxin 3	Control
28	33	18	11
23	36	21	14
14	34	20	11
27	29	22	16
	31	24	
	34		

```
> CIs <- sapply(Observations, get_bound, k=k, alpha=0.05)
> CIs
          [,1]     [,2]     [,3]     [,4]
[1,]  5.807657 28.97077 16.68534  6.396238
[2,] 40.192343 36.69589 25.31466 19.603762
```

```
library(ggplot2)
centers <- sapply(Observations, mean)
CI_df <- data.frame(lab = group_ind, center = centers,
                    lower = CIs[1,], upper = CIs[2,])
ggplot(CI_df, aes(x = lab, y = center)) + geom_point() +
  geom_errorbar(width = 0.1, aes(ymin = lower, ymax = upper)) +
  geom_hline(yintercept = mean(unlist(Observations)), col = 2)
```



Tomorrow ...

- Compute Tukey Honest Significant Differences (HSD);
- The Kruskal-Wallis test;
- Two-way ANOVA.