3703-3J04 Probability & Statistics for (Civil) Engineering

C01 - Lecture 35

Last time Single-Factor Experiments & ANOVA

- 1 factor investigated at a different levels (treatments)
- n observations for each treatment

- Model each observation Y_{ij} as $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$
  - μ: overall mean
  - τ_i: ith "treatment effect"
  - ε_{ij}: random error, assume Normal
- Randomize experiment; then $\varepsilon_{ij} \sim N(0, \sigma^2)$ i.e. all variances same

$i = 1, \ldots, a$

$j = 1, \ldots, n$

Recall mean of ith treatment $\mu_i = \mu + \tau_i$

Want to test $H_0 : \mu_1 = \mu_2 = \ldots = \mu_a$

H_1 : at least one pair (i,k) has $\mu_i \neq \mu_k$.

We can estimate $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ with estimators for μ, τ_i and ε_{ij}.

Notation

The dot always means "sum over the variable that the dot replaces"

$$Y_{i\bullet} = \sum_{j=1}^{n} Y_{ij} = \text{total under ith treatment}$$

(here j)

$$\bar{Y}_{i\bullet} = \frac{1}{n} Y_{i\bullet} = \text{sample mean under ith treatment}$$

$$Y_{\bullet\bullet} = \sum_{i=1}^{a} Y_{i\bullet} = \sum_{i=1}^{a}\sum_{j=1}^{n} Y_{ij} = \text{grand total}$$

$N = na$ = total # of observations

$\bar{Y}_{..} = \frac{1}{N} Y_{..}$ = grand mean

$$\left[S_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i.})^2 = \text{sample variance under } i\text{th treatment} \right]$$

We have $\hat{\mu} = \bar{Y}_{..}$ (estimator of μ)

$\hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{..}$ (estimator of τ_i)

Residual → $e_{ij} = \hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_{i.}$ (estimator of ε_{ij})
(Error within group)

Notice we have $Y_{ij} = \bar{Y}_{..} + (\bar{Y}_{i.} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i.})$

$$\rightarrow \quad Y_{ij} - \bar{Y}_{..} = (\bar{Y}_{i.} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i.})$$

It can be shown that squaring both sides & summing over i & j gives:

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{..})^2}_{= SS_T} = \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n} (\bar{Y}_{i.} - \bar{Y}_{..})^2}_{= SS_{Treatments}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i.})^2}_{= SS_E}$$

= Total variation in data

= "Between groups sum of squares"

Sorry, this was a MISTAKE

The software packages call $SS_{Treat.}$ the "between groups s.o.s."

= Variation due to differences in the treatment means (comes from $\hat{\tau}_i$)

= Random variation (comes from e_{ij})

"Within groups sum of squares"

↳ To find SS_E = SS_T − $SS_{Treat.}$

↳ Shortcut formula:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n} Y_{ij}^2 - N\bar{Y}_{..}^2$$

Notice this $= \frac{Y_{..}^2}{N}$

as found in textbook!

↳ Shortcut formula:

$$\sum_{i=1}^{a}\sum_{j=1}^{n} \bar{Y}_{i.}^2 - N\bar{Y}_{..}^2 \quad (= \frac{Y_{..}^2}{N})$$

$$= \sum_{i=1}^{a} \frac{Y_{i.}^2}{n} - N\bar{Y}_{..}^2$$

$$E(SS_T) = E(SS_{Treatments}) + E(SS_E)$$

$$E(SS_{Treatments}) = (a-1)\sigma^2 + n\sum_{i=1}^{a} \tau_i^2$$

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n} Y_{ij}^2 - \sum_{i=1}^{a} \frac{Y_{i.}^2}{n}$$

$$E(SS_E) = a(n-1)\sigma^2$$

The coefficient of σ^2 in each is # of degrees of freedom.

We define the "mean squares" for treatments/error by:

$$MS_{Treatments} = SS_{Treatments} / (a-1)$$

$$MS_E = SS_E / a(n-1).$$

Notice $E(MS_E) = \dfrac{E(SS_E)}{a(n-1)} = \dfrac{a(n-1)\sigma^2}{a(n-1)} = \sigma^2$

So MS_E is always an unbiased estimator for σ^2.

But $E(MS_{Treatments}) = \dfrac{E(SS_{Treat.})}{a-1}$

$$= \frac{(a-1)\sigma^2 + n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

$$= \sigma^2 + \frac{n}{a-1}\sum_{i=1}^{a}\tau_i^2$$

so $MS_{Treatments}$ is unbiased for σ^2 only when $\sum_{i=1}^{a}\tau_i^2 = 0$ i.e. only if H_0 true.

H_0 says $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_a$

i.e. $\mu + \tau_1 = \mu + \tau_2 = \ldots = \mu + \tau_a$

i.e. $\tau_1 = \tau_2 = \ldots = \tau_a$.

Notice also $\mu = \dfrac{1}{a}\sum_{i=1}^{a}\mu_i$, the mean of means

$$= \frac{1}{a}\sum(\mu + \tau_i)$$

$$\mu = \mu + \frac{1}{a}\sum_{i=1}^{a}\tau_i \rightarrow \sum_{i=1}^{a}\tau_i = 0$$

(always)

So under H_0: $\tau_1 = \tau_2 = \ldots = \tau_a = 0$

So H_1 says some (in fact at least 2) $\tau_i \neq 0$

& if H_1 is true then $MS_{Treatments}$ overestimates σ^2 on average.

The question is, does it overshoot by a significant amount?

The test statistic $F_0 = \dfrac{MS_{Treatments}}{MS_E}$

$\cong 1$ if H_0 true

$>> 1$ if H_1 true

So the F-test tells us when F_0 being >1 is significantly bigger than 1.

All we now need to test is to know distribution:

F_0 has F-distribution with $(a-1)$ d.o.f. in numerator & $a(n-1)$ d.o.f. in denominator.

Rule: Reject H_0 at level α if

$f_0 > f_{\alpha, a-1, a(n-1)}$

f_0: computed value of F_0

$f_{\alpha, a-1, a(n-1)}$: value from the F_α-table

Example

Source: A. Parenti, L. Guerrini, P. Masella, S. Spinelli, L. Calamai, P. Spugnoli (2014). "Comparison of Espresso Coffee Brewing Techniques," Journal of Food Engineering, Vol. 121, pp. 112-117.

Description: Comparison of foam index (Y, in %) for 3 methods of brewing espresso
Method 1 = Bar Machine (BM),
Method 2 = Hyper-Espresso Method (HIP),
Method 3 = I-Espresso System (IT).

a=3

9 replicates/treatment. **n=9**

Variables: foamIndx (response), method (factor)

method	1		2		3	
	y_{1j}		y_{2j}		y_{3j}	
foamIndx	36.64	1342.4896	70.84	5018.31	56.19	3157.32
	39.65	1572.1225	46.68	2179.02	36.67	1344.69
	37.74	1424.3076	73.19	5356.78	35.35	1249.62
	35.96	1293.1216	57.78	3338.53	40.11	1608.81
	38.52	1483.7904	48.61	2362.93	33.52	1123.59
	21.02	441.8404	72.77	5295.47	37.12	1377.89
	24.81	615.5361	65.04	4230.2	37.33	1393.53
	34.18	1168.2724	62.53	3910	32.68	1067.98
	23.08	532.6864	54.26	2944.15	48.33	2335.79
$y_{1.}$	**291.6**		**$y_{2.}$**	**551.7**	**$y_{3.}$**	**357.3**
$\bar{y}_{1.}$	**=291.6/9**		**$\bar{y}_{2.}$**	**=551.7/9**	**$\bar{y}_{3.}$**	**=357.3/9**
	32.4			**61.3**		**39.7**
$(y_{1.})^2$	**85030.56**		**$(y_{2.})^2$**	**304373**	**$(y_{3.})^2$**	**127663**

$y_{..}$ **1200.6**
$\bar{y}_{..}$ **=1200.6/27**
44.4666667
$(\bar{y}_{..})^2$ **1977.28444**
$\Sigma(y_{ij})^2 =$ **59168.7792**

SS_T **4065.18**
$SS_{Treatments}$ **1716.9192**

Totals within each "treatment" (method/group)

Mean within each treatment

Method	1	2	3
	36.64	70.84	56.19
	39.65	46.68	36.67
	37.74	73.19	35.35
	35.96	57.78	40.11
	38.52	48.61	33.52
	21.02	72.77	37.12
	24.81	65.04	37.33
	34.18	62.53	32.68
	23.08	54.26	48.33

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Method 1	9	291.6	32.4	53.29087
Method 2	9	551.7	61.3	102.0222
Method 3	9	357.3	39.7	59.30182

This is called an ANOVA table.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
(Treatments) Between Groups	4065.18	2	2032.59	28.41261	4.69864E-07	3.402826
(Error) Within Groups	1716.919	24	71.5383			
Total	5782.099	26				

$a-1$

f_0

$a(n-1) = 3(9-1)$

This is the value for $\alpha = 0.05$

So do we reject $H_0: \mu_1 = \mu_2 = \mu_3$?

$F = f_0 = 28.41$ is WAY bigger than $F_{crit} = f_{0.05,24,2}$

so reject H_0 at level $\alpha = 0.05$.

i.e. $f_{0.05,24,2} = 3.402826$ → see f-table