Multiple Means Testing

Recall the setup from the last example:

$X_{1i} \sim N(\mu_1, \sigma^2)$ $\quad X_{2i} \sim N(\mu_2, \sigma^2)$ $\quad X_{3i} \sim N(\mu_3, \sigma^2)$

where our interest is in

$H_0: \mu_1 = \mu_2 = \mu_3$ vs. H_A: at least one pair unequal

We found a likelihood ratio test by testing

$\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} < k$ for some k.

ANOVA (= ANalysis Of VAriance) is all about a systematic approach to this likelihood ratio test (and generalizations)

ANOVA and the T-test fit into a sequence of <u>means tests</u>:

- one-sample $\quad H_0: \mu = \mu_0$
- two-sample $\quad H_0: \mu_1 = \mu_2$
- many-sample (ANOVA) $\quad H_0: \mu_1 = \dots = \mu_k$
- continuous spectrum of samples (regression)

Variances plot

Notation

Since ANOVA is all about sample variances, there will be <u>many</u> sums of squares involved.

We write

$Y_{ij} \sim N(\mu_i, \sigma^2)$ $\quad$ n_i samples from each of k groups

$$\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \qquad \bar{Y} = \bar{Y}_{\cdot\cdot} = \frac{1}{\sum n_i} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

Total Sum of Squares $\quad TSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$

Sum of Squares for Errors $\quad SSE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$

Sum of Squares for Treatments $\quad SST = \sum_{i=1}^{k} (\bar{Y}_{i\cdot} - \bar{Y})^2$

Mean square for Errors $\quad MSE = SSE/DoF$

Mean square for Treatments $\quad MST = SST/DoF$

Each of $\frac{1}{n-1} TSS$, MSE, MST is an unbiased estimator of σ^2.

Theorem

$$TSS = SST + SSE$$

Proof $TSS = \sum_i \sum_j (Y_{ij} - \bar{Y})^2 = \sum_i \sum_j \left((Y_{ij} - \bar{Y}_{i\cdot}) + (\bar{Y}_{i\cdot} - \bar{Y})\right)^2$

$$= \sum_i \sum_j \left[(Y_{ij} - \bar{Y}_{i\cdot})^2 + 2(Y_{ij} - \bar{Y}_{i\cdot})(\bar{Y}_{i\cdot} - \bar{Y}) + (\bar{Y}_{i\cdot} - \bar{Y})^2\right]$$

Notice that $\sum_j (Y_{ij} - \bar{Y}_{i\cdot}) = \sum_j Y_{ij} - n_i \bar{Y}_{i\cdot} = n_i \bar{Y}_{i\cdot} - n_i \bar{Y}_{i\cdot} = 0$

So $\sum_j 2(Y_{ij} - \bar{Y}_{i\cdot})(\bar{Y}_{i\cdot} - \bar{Y}) = 2(\bar{Y}_{i\cdot} - \bar{Y}) \sum (Y_{ij} - \bar{Y}_{i\cdot}) = 0.$

The remainder is $\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y})^2.$ ∎

The ANOVA F-test

Recall $SSE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$.

Since the sample variance of the i^{th} sample on its own is $\frac{1}{n_i - 1} \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$, it follows that SSE is a pooled sum of squares

$$SSE = \sum_i (n_i - 1) S_i^2 \quad \text{where} \quad S_i^2 \text{ is the } i^{th} \text{ sample variance.}$$

Theorem If $U \sim \chi^2(n)$ and $V \sim \chi^2(m)$, independent, then $U+V \sim \chi^2(n+m)$.

Proof sketch If $U = \sum_{i=1}^{n} Z_i$ and $V = \sum_{j=1}^{m} W_j$, with $Z_i, W_j \sim N(0,1)$, then $U+V$ is a sum of squares of $n+m$ iid $N(0,1)$ variables. ∎

Since each $(n_i - 1) S_i^2 / \sigma^2 \sim \chi^2(n_i - 1)$, it follows that

$$\frac{SSE}{\sigma^2} = \sum_i (n_i - 1) \frac{S_i^2}{\sigma^2} \sim \chi^2\left(\sum_i (n_i - 1)\right) = \chi^2(n-k).$$

Under the null hypothesis $\mu_1 = \cdots = \mu_k$, all the Y_{ij} are iid.
Hence, $TSS = \sum_i \sum_j (Y_{ij} - \bar{Y})^2 = (n-1) \cdot S^2$ for $n = \sum n_i$ and S^2 the sample variance of the union of all the Y_{ij}'s.

Hence $\frac{TSS}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

In order to understand SST, we will need a theorem on subtracting χ^2 variables:

Theorem If $U \sim \chi^2(m)$, $W = U+V \sim \chi^2(n+m)$ then $V \sim \chi^2(n)$.

Proof The moment generating function for W is:

$$(1-2t)^{-(n+m)/2} = \mathbb{E}\left[e^{t(U+V)}\right] = \mathbb{E}\left[e^{tU} \cdot e^{tV}\right] =$$

$$= \mathbb{E}\left[e^{tU}\right] \cdot \mathbb{E}\left[e^{tV}\right] = (1-2t)^{-m/2} \cdot MGF_V(t)$$

because independent

So $MGF_V(t) = (1-2t)^{-(n+m)/2} / (1-2t)^{-m/2} = (1-2t)^{-n/2}$

Hence $V \sim \chi^2(n)$. ∎

Since $TSS = SSE + SST$, $\frac{TSS}{\sigma^2} \sim \chi^2(n-1)$ and $\frac{SSE}{\sigma^2} \sim \chi^2(n-k)$ it follows that $\frac{SST}{\sigma^2} \sim \chi^2((n-1)-(n-k)) = \chi^2(k-1)$.

We define $MSE = SSE/(n-k)$ and $MST = SST/(k-1)$.

Then

$$F = \frac{\frac{SST}{\sigma^2}/(k-1)}{\frac{SSE}{\sigma^2}/(n-k)} = \frac{SST/(k-1)}{SSE/(n-k)} \sim F^{k-1}_{n-k}$$

The ANOVA table

Since there are so many components to the F-statistic calculation, it is often helpful to organize them in a table:

Source	DoF	SS	MS	F
Treatments	$k-1$	SST	MST	MST/MSE
Error	$n-k$	SSE	MSE	
Total	$n-1$	TSS		

Example (13.11)

Data was given:

	G1	G2	G3
n	14	14	14
$\bar{x}_{i\cdot}$	0.93	1.21	0.92
std.err.	0.04	0.03	0.04

Since std.err $= s/\sqrt{n}$, we get

s_i	0.15	0.11	0.15
s_i^2	0.022	0.012	0.022

Now, $SSE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 = \sum_i (n_i - 1) S_i^2$

and $SST = \sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y})^2 = \sum_i n_i (\bar{Y}_{i\cdot} - \bar{Y})^2$

where $\bar{Y} = \frac{1}{n} \sum n_i \bar{Y}_{i\cdot}$

Source	DoF	SS	MS	F	P
T	2	0.7588	0.3794	19.829	$1.1 \cdot 10^{-6}$
E	39	0.7462	0.0191		

Estimation

MSE is an <u>unbiased</u> pooled estimator of σ^2.
It produces better estimates than would the sample variance from any one group in isolation.

We get confidence intervals:

$$\mu_i \in \bar{Y}_{i\cdot} \pm t_{\alpha/2} \, S/\sqrt{n_i}$$

$$\mu_i - \mu_j \in \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm t_{\alpha/2} \, S\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where $S = \sqrt{MSE}$ and $t_\alpha = F^{-1}_{t(n-k)}(1-\alpha)$.

Errors using this compound if you repeat CI calculations for other group means or group mean differences.

One very common method to deal with this is the <u>Bonferroni correction</u>:
Suppose we are seeking confidence intervals $I_1, \ldots, I_m$ for parameters $\vartheta_1, \ldots, \vartheta_m$ such that

$$\mathbb{P}(\vartheta_i \in I_i \text{ for all } i) \geq 1-\alpha.$$

Let's talk a bit about events:

$$\overline{A_1 \cap A_2 \cap \cdots \cap A_m} = \bar{A}_1 \cup \bar{A}_2 \cup \cdots \cup \bar{A}_m$$

So if the additive law applies, then

$$\begin{aligned}\mathbb{P}(A_1 \cap \cdots \cap A_m) &= 1 - \mathbb{P}(\bar{A}_1 \cup \cdots \cup \bar{A}_m) \\ &\geq 1 - \sum \mathbb{P}(\bar{A}_i) \\ &= 1 - \sum \alpha_j\end{aligned}$$

So for our confidence interval, if each is an $(1-\alpha)$-CI, then the <u>joint confidence level</u> could be as small as $1 - m\alpha$.

Bonferroni's method: use $\alpha' = \alpha/m$ for each simultaneous interval.

Additivity is usually not applicable and Bonferroni is known to be too conservative (ie rejects too rarely). Better methods have been proposed — eg by Holm and by Hochberg — but are out of scope for this course.