Qualche riga in piú sul F-Test.

Consideriamo una serie di **K** esperimenti. In ognuno dei quali sono state fatte un certo numero (m_k) di osservazioni di una variabile aleatoria X ($x_{k,i}$). Ogni esperimento permette di ottenere un valor medio ($\bar{X}_k$) e la stima campionaria della varianza (s_k^2) per il campione dato. L'ipotesi che si vuole verificare é se i singoli esperimenti possono essere considerati come campionamenti della stessa "popolazione", quindi se i valori medi e le varianze trovate possono essere considerati come stima del valore aspettato e della varianza della popolazione globale.

	Exp_1	Exp_2	. . .	Exp_K
	m_1	m_2		m_K
	$x_{1,1}$	$x_{2,1}$		$x_{K,1}$
	$x_{1,2}$	$x_{2,2}$		$x_{K,2}$
	$x_{1,3}$	$x_{2,3}$		$x_{K,3}$
	...	...		
	x_{1,m_1}	x_{2,m_2}		x_{K,m_2}
medie:	$\bar{X}_1$	$\bar{X}_2$		$\bar{X}_K$
dev.st:	s_1	s_2		s_K

Dato un insieme di K esperimenti, ognuno numerositá m_k, assumiamo che tutte le misure vengono dalla stessa popolazione. In tal caso il valore medio, calcolato su tutte le osservazioni ($x_{j,k}$) é:

$$\bar{X} = \frac{\sum_{i=1}^{m_1} x_i + \sum_{i=1}^{m_2} x_i + ... + \sum_{i=1}^{m_K}}{m_1 + m_2 + ... + m_k} = \frac{\sum_{k=1}^{K} m_k \bar{X}_k}{\sum_{k=1}^{K} m_k} = \frac{\sum_{k=1}^{K} m_k \bar{X}_k}{m}$$

con $m = m_1 + m_2 + ...m_k$. In pratica é una media pesata: il valor medio dell'esperimento k-esimo é pesato per la sua molteplicitá m_k. Nota: la media pesata é eguale alla media su tutte le osservazioni di tutti i campioni. Al contrario la media delle medie non é uguale alla media su tutte le osservazioni.

Prima di andare avanti notiamo che la varianza campionaria su un insieme di osservazioni é:

$$\begin{aligned}
s^2 &= \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{X})^2 = \\
&= \frac{1}{m-1}\sum_{i=1}^{m}(x_i^2 + \bar{X}^2 - 2x_i\bar{X}) = \frac{1}{m-1}\sum_{i=1}^{m}(x_i^2 - x_i\bar{X}) = \\
&= \frac{1}{m-1}\left(\sum_{i=1}^{m} x_i^2 - m\bar{X}\frac{1}{m}\sum_{i=1}^{m} x_i\right) = \frac{1}{m-1}\left(\sum_{i=1}^{m} x_i^2 - m\bar{X}^2\right) = \\
&= \frac{m}{m-1}\left(\frac{1}{m}\sum_{i=1}^{m} x_i^2 - \bar{X}^2\right) = \\
&= \frac{m}{m-1}\left(\overline{X^2} - \bar{X}^2\right)
\end{aligned}$$

Quindi la varianza campionaria puó essere calcolata come somma del quadrato degli scarti diviso m-1, oppure come differenza tra la media dei quadrati delle osservazioni e il quadrato della media, normalizzata per $m/m-1$. Questa espressione puó essere utile per semplificare le espressioni: ad esempio, avendo a che fare con K esperimenti ognuno con un valor medio $\bar{X}_k$, la migliore stima della media é $\bar{X}$, come visto sopra (e non la media delle medie, infatti é intuitivo che campioni con una numerositá maggiore forniscano una migliore stima dell valore aspettato vero della popolazione), l'errore standard sulla media si puó stimare come:

$$\begin{aligned} s^2_{\bar{X}} &= \frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{X}_k-\bar{X}\right)^2 = \frac{1}{K-1}\left(\sum_{k=1}^{K}\left(\bar{X}_k^2+\bar{X}^2-2\bar{X}\bar{X}_k\right)\right) = \\ &= \frac{1}{K-1}\left(\sum_{k=1}^{K}\bar{X}_k^2+K\bar{X}^2-2\bar{X}\sum_{k=1}^{K}\bar{X}_k\right) = \\ &= \frac{1}{K-1}\left(\sum_{k=1}^{K}\bar{X}_k^2+K\bar{X}^2-2K\bar{X}^2\right) = \\ &= \frac{1}{K-1}\left(\sum_{k=1}^{K}\bar{X}_k^2-K\bar{X}^2\right) = \\ &= \frac{K}{K-1}\left(\overline{\bar{X}^2}-\bar{X}^2\right) \end{aligned}$$

dove $\overline{\bar{X}^2}$ é la media dei quadrati delle medie campionarie di ogni esperimento, invece $\bar{X}^2$ é il quadrato della media stimata utilizzando tutti i dati di tutti gli esperimenti.

Se si puó vuole stimare la varianza di tutte le osservazioni possiamo utilizzare due approcci: nel primo caso (s^2_{tra}) assumiamo che, nell'ipotesi da verificare che tutti i dati vengano dalla stessa popolazione, tutti gli m_k valori nell'esperimento k-esimo abbiano uno stesso valore medio atteso, la cui miglior stima é appunto $\bar{X}_k$. Quindi stimiamo s^2_{tra} come media dei quadrati degli scarti tra media dell'esperimento k-esimo e $\bar{X}_k$. Dal momento che esperimenti con maggiore numerositá sono probabilmente piú precisi, pesiamo i quadrati degli scarti per la numerositá del campione k-esimo:

$$\begin{aligned} s^2_{tra} &= \frac{1}{K-1}\sum_{k=1}^{K}m_k\left(\bar{X}_k-\bar{X}\right)^2 = \frac{1}{K-1}\left(\sum_{k=1}^{K}m_k\left(\bar{X}_k^2+\bar{X}^2-2\bar{X}\bar{X}_k\right)\right) = \\ &= \frac{1}{K-1}\left(\sum_{k=1}^{K}m_k\bar{X}_k^2+\bar{X}^2\sum_{k=1}^{K}m_k-2\bar{X}\sum_{k=1}^{K}m_k\bar{X}_k\right) = \\ &= \frac{1}{K-1}\left(\sum_{k=1}^{K}m_k\bar{X}_k^2+m\bar{X}^2-2m\bar{X}^2\right) = \\ &= \frac{1}{K-1}\left(\sum_{k=1}^{K}m_k\bar{X}_k^2-m\bar{X}^2\right) \end{aligned}$$

Notiamo che, se le numerositá dei campioni sono le stesse ($m_1=m_2=...m_k=m_o$), si ha:

$$s^2_{tra} = \frac{m_o}{K-1}\sum_{k=1}^{K}\left(\bar{X}_k-\bar{X}\right)^2 = \frac{1}{K-1}\left(m_o\sum_{k=1}^{K}\bar{X}_k^2-m_oK\bar{X}^2\right)$$

$$= m_o \frac{K}{K-1}\left(\frac{1}{K}\sum_{k=1}^{K}\bar{X}_k^2 - \bar{X}^2\right) = m_o \frac{K}{K-1}\left(\overline{\bar{X}^2} - \bar{X}^2\right)$$

dove $\overline{\bar{X}^2}$ é la media dei quadrati delle medie campionarie di ogni singolo esperimento mentre $\bar{X}^2$ é la media stimata utilizzando tutti i dati di tutti gli esperimenti. Per campioni della stessa numerositá si ha:$s^2_{tra} = m_o s^2_{\bar{x}}$ che giustifica la relazione tra stima della varianza e errore standard della media: $s^2_{\bar{x}} = s^2/\sqrt{m_o}$

Un secondo modo per stimare la varianza delle osservazioni é come media pesata delle varianze campionarie, in cui il peso é il numero di gradi di libertá per ognuna di esse:

$$s^2_{intra} = \frac{\sum_{k=1}^{K}(m_k-1)s_k^2}{\sum_{k=1}^{K}(m_k-1)} = \frac{\sum_{k=1}^{K}(m_k-1)s_k^2}{m-K}$$

che, per campioni della stessa numerositá, diventa:

$$s^2_{intra} = \frac{1}{K}\sum_{k=1}^{K}s_k^2$$

A volte s^2_{intra} si trova anche scritta con un formalismo un pó piú complesso:

$$\begin{aligned} s^2_{intra} &= \frac{\sum_{k=1}^{K}(m_k-1)s_k^2}{\sum_{k=1}^{K}(m_k-1)} = \frac{1}{m-K}\sum_{k=1}^{K}(m_k-1)\frac{m_k}{m_k-1}\left(\overline{X_k^2} - \bar{X}_k^2\right) = \\ &= \frac{1}{m-K}\sum_{k=1}^{K} m_k\left(\frac{1}{m_k}\sum_{i=1}^{m_k}x_{k,i}^2 - \bar{X}_k^2\right) = \\ &= \frac{\sum_{k=1}^{K}\sum_{i=1}^{m_k}x_{k,i}^2 - \sum_{k=1}^{K}m_k\bar{X}_k^2}{m-K} \end{aligned}$$

Se i campioni vengono effettivamente dalla stessa popolazione (ipotesi da verificare), allora il rapporto: $F = s^2_{tra}/s^2_{intra} \sim 1$. In generale si puó dimostrare che $s^2_{intra} < s^2_{tra}$ e che ció é tanto piú vero quanto piú sono diverse le popolazioni d'origine dei campioni.

La funzione F é una variabile aleatoria che segue una distribuzione nota come distribuzione F, la cui forma dipende dai gradi di liertá al numeratore ($\nu_1 = K-1$) e al denominatore ($\nu_2 = m-K$). Questa funzione si trova tabulata in funzione di ν_1, ν_2 e del livello di confidenza: $F(\nu_1, \nu_2, \alpha)$. Il valore riportato nelle tabelle indica il valore critico F_c con il seguente significato: nell'ipotesi che i campioni siano tutti provenienti da una stessa popolazione la probabilitá di osservare valori $F > F_c$ é minore di α mentre la probabilitá di osservare valori di $F < F_c$ é maggiore di $1-\alpha$. Quindi se si ottiene un valore $F < F_c$ possiamo dire che l'ipotesi é verificata con un margine di errore (rischio) pari ad α (o che é la stessa cosa l'ipotesi é verificata con probabilitá $1-\alpha$). Quindi posso assumere che le differenze eventualmente osservate nei diversi campioni on sono significative da un punto di vista statistico.

Al contrario se si ottiene un valore $F > F_c$ possiamo dire che l'ipotesi non é verificata, quindi i campioni presentano diversitá statisticamente significative, con un margine di errore (rischio) pari ad α (o che é la stessa cosa l'ipotesi non é verificata con una confidenza $1-\alpha$).

In Excel la funzione `INV.F(n1,n2,alpha)` restituisce il valore F_c per un livello di confidenza α.