Note 5 (Some notes about the F-test in ANOVA)

- Connection between the 2-sample (unpaired) t-test and the F-test in ANOVA (I samples, $I \geqslant 2$): the 2-sample t-test is a special case of the F-test where only two groups are being compared ($I = 2$) because J_1, Y_{1j}, J_2, Y_{2j}
  - test statistic (**recall**: LN, CH11, p.13, t-test for n X_i's and m Y_j's)
    - $SS_B/(I-1) = SS_B$ ($\bar{Y}_{..}$, $\bar{Y}_{1.}$, $\bar{Y}_{2.}$)

$$= \underset{J_1}{n}\left[\overline{X} - \left(\tfrac{n}{m+n}\overline{X} + \tfrac{m}{m+n}\overline{Y}\right)\right]^2 + \underset{J_2}{m}\left[\overline{Y} - \left(\tfrac{n}{m+n}\overline{X} + \tfrac{m}{m+n}\overline{Y}\right)\right]^2$$

$$= n\frac{m^2}{(m+n)^2}(\overline{X}-\overline{Y})^2 + m\frac{n^2}{(m+n)^2}(\overline{X}-\overline{Y})^2 = \frac{mn}{m+n}(\overline{X}-\overline{Y})^2$$

    - $SS_W/(N-I) = SS_W/[(m-1)+(n-1)] = s_p^2$ (LN, CH11, P.7, Definition 1)

2-sided $|T|>c$ becomes 1-sided $F>c'$

$$\circledast\ F = \frac{SS_B/(I-1)}{SS_W/(N-I)} = \frac{(\overline{X}-\overline{Y})^2}{s_p^2\left(\frac{1}{n}+\frac{1}{m}\right)} = T^2$$

indep. $t_d = \dfrac{N(0,1)}{\sqrt{\chi^2_d/d}}$, $t_d^2 = \dfrac{[N(0,1)]^2}{\chi^2_d/d}$, $\sim \chi^2_1/1$

$d = N-I = m+n-2$

⊖ null distribution: if $Z \sim t_d$, then $Z^2 \sim F_{1,d}$

- Under the model ($\ast$) in LNp.5, the F-test in ANOVA is equivalent to the likelihood ratio test (exercise, the proof is similar to what presented in LN, CH11, p.15-18, for the case of two independent samples).

• Under $\Omega = H_0 \cup H_A$, MLEs: $\hat{\mu}_{i,\Omega} = \bar{Y}_{i.}$, $i=1,2,\cdots,I$, $\hat{\sigma}^2_\Omega = SS_W/N$
• Under $\omega = H_0$, MLEs: $\hat{\mu}_{1,\omega} = \cdots = \hat{\mu}_{I,\omega} = \bar{Y}_{..}$, $\hat{\sigma}^2_\omega = SS_{TOT}/N$

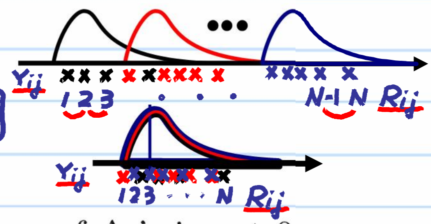
- **A nonparametric method --- the Kruskal-Wallis test** ← Why need it? Check Note 6, LN, CH11, p.25-26.
  - Consider the model ($\otimes$) in LNp.1, and further assume that
    - ⊖ Ω is the collection of *all* continuous distributions $\Rightarrow$ $\dim(\Omega) = \infty$
    - ⊖ $F_1, \ldots, F_I$ have the same shape, i.e.,

$$F_1 = F(x-\Delta_1), \ldots, F_I = F(x-\Delta_I),$$

where $F \in \Omega$.

Why need this assumption? Check Q.7 (LN, CH11, p35)

$F_1, \cdots, F_I$ have same variance



  - Test the null and alternative hypotheses:

$$H_0: \Delta_1 = \cdots = \Delta_I = 0 \quad \text{vs.} \quad H_A: \text{at least one of } \Delta_i\text{'s is not } 0$$

★ ⊙ Under H_0, all Y_{ij}'s $\sim$ i.i.d. F. → Under H_0, the distribution of ranks is irrelevant to F.

  - **Recall**. The sample size of the ith sample is J_i, $i = 1, \ldots, I$, and $N = J_1 + \cdots + J_I$ is the number of all observations.

$Y_{ij} \to R_{ij}$ cf.

Theorem 9 (Kruskal-Wallis test)

use ranks, rather than raw data, to do analysis

⊙ Let R_{ij}'s be the *ranks* of Y_{ij}'s in the *combined* (*pooled*) sample.

- Define

weighted average of $\bar{R}_{1.}, \cdots, \bar{R}_{I.}$

$\bar{Y}_{i.}$ cf. ⊖ $\bar{R}_{i.} = \dfrac{1}{J_i}\sum_{j=1}^{J_i} R_{ij}$: average rank in the ith group

$\bar{Y}_{..}$ cf. ⊖

$$\bar{R}_{..} = \frac{1}{N}\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J_i} R_{ij}}_{1+2+\cdots+N} = \frac{J_1\bar{R}_{1.} + \cdots + J_I\bar{R}_{I.}}{J_1 + \cdots + J_I} = \frac{1}{N}\overbrace{\frac{N(N+1)}{2}}^{=1+\cdots+N} = \frac{N+1}{2}$$

a constant, not r.v.

- Define

$Var(Z)=E(Z^2)-[E(Z)]^2$, $Z=\bar{R}_{i\cdot}$ with prob. J_i/N

$$SS_B=\sum_{i=1}^{I} J_i(\bar{R}_{i\cdot}-\bar{R}_{\cdot\cdot})^2=\left(\sum_{i=1}^{I} J_i\bar{R}_{i\cdot}^2\right)-N\bar{R}_{\cdot\cdot}^2=\left(\sum_{i=1}^{I} J_i\bar{R}_{i\cdot}^2\right)-\frac{N(N+1)^2}{4}$$

Between

which measures *dispersion* of $\bar{R}_{i\cdot}$'s. — large, if Δ_i's very different; small, if Δ_i's about the same

- Test statistic K

$$F\propto\frac{SS_B}{SS_W}\ \overset{cf.}{\longleftrightarrow}\ K=\frac{12}{N(N+1)}SS_B=\frac{12}{N(N+1)}\left(\sum_{i=1}^{I} J_i\bar{R}_{i\cdot}^2\right)-3(N+1)$$

(**Note**. SS_B can be found by running R_{ij}'s through an ANOVA program)

– Q: Why is SS_B divided by $\frac{N(N+1)}{12}$ in K? Define

$=1^2+2^2+\cdots+N^2=\frac{N(N+1)(2N+1)}{6}$

$Var(Z)=E(Z^2)-[E(Z)]^2$, $Z=R_{ij}$ with prob. $1/N$

a constant, not r.v.

$$SS_{TOT}=\sum_{i=1}^{I}\sum_{j=1}^{J_i}(R_{ij}-\bar{R}_{\cdot\cdot})^2=\left(\sum_{i=1}^{I}\sum_{j=1}^{J_i}R_{ij}^2\right)-N\bar{R}_{\cdot\cdot}^2=\frac{(N-1)N(N+1)}{12}$$

Then,

$\sigma^2 \leftarrow e\,(H_0 \cup H_A)$ $\frac{SS_W}{N-I}$ cf.

$\sigma_R^2 \leftarrow e$ (under H_0)

a constant

$$K=\frac{SS_B}{SS_{TOT}/(N-1)}=\frac{SS_B}{(SS_B+SS_W)/(N-1)}=\frac{1}{\left(1+\frac{SS_W}{SS_B}\right)/(N-1)}$$

$\Rightarrow$ $F\propto\frac{SS_B}{SS_W}\uparrow$ $\Leftrightarrow$ $\frac{SS_B}{SS_{TOT}}\uparrow$ $\Leftrightarrow$ $K\uparrow$

Note. $SS_{TOT}=SS_B+SS_W$ (LNp.7) still holds for R_{ij}'s.

a constant $-SS_B$

– Q: Why is no SS_W in K?

Q: Why ANOVA use indep. $\frac{SS_B}{SS_W}$, not $\frac{SS_B}{SS_{TOT}}$ not indep.

– Data with large values of K are more extreme, i.e., provide stronger evidence against H_0.