Analysis of variance

In its simplest form we assume that we have I different populations. Each population has a population mean μ_i and a population variance σ_i^2. Suppose you have a sample of size n_i from each of the I populations. Assume the sample values $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ from the ith populations are i.i.d. normal with $E(Y_{ij}) = \mu_i$ and $\text{var}(Y_{ij}) = \sigma_i^2$. We wish to test

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_I \qquad \text{versus} \qquad H_1\colon \text{not all } \mu_i \text{ are equal.}$$

We will assume that $\sigma_1^2 = \cdots = \sigma_I^2$.

a. The figure below shows two possibilities for histograms of the samples.

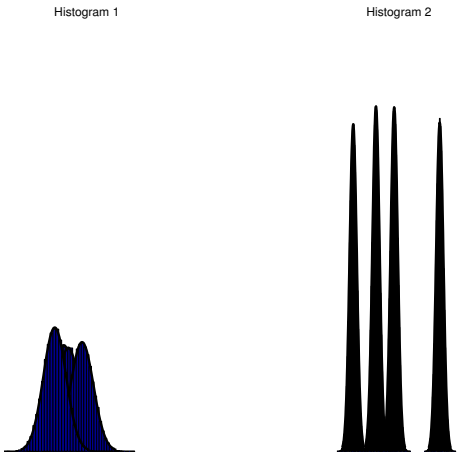

Figure 1 Two possible histograms for $I = 4$.

In which case would you judge that the population means are different. Why? Explain.

b. In analysis of variance we compare the dispersion of sample means (variance bewteen groups) with the dispersion within groups (variance within). Fro simplicity assume $n_1 = n_2 = \cdots = n_I = J$. Denote

$$Y_{i\cdot} = \frac{1}{J} \sum_{i=1}^{J} Y_{ij}\,.$$

If H_0 holds the variance of sample means is $\sigma/\sqrt{J}$. If the sample means are "too" dispersed compared to $\sigma/\sqrt{J}$ we reject H_0. But some theoretical work is needed.

(i) Naj bodo $Y_1, \ldots, Y_J$ med sabo neodvisne normalne slučajne spremenljivke z enako porazdelitvijo. Označimo z μ skupno matematično upanje in z σ^2 skupno varianco. Pokažite, da je povprečje $\bar{Y}$ neodvisno od vektorja $(Y_1 - \bar{Y}, \ldots, Y_J - \bar{Y})$. Ker je $(\bar{Y}, Y_1 - \bar{Y}, \ldots, Y_J - \bar{Y})$ večrazsežen normalen vektor, je dovolj pokazati, da sta $\bar{Y}$ in $(Y_1 - \bar{Y}, \ldots, Y_J - \bar{Y})$ nekorelirana. Upoštevajte še, da je

$$(Y_1 - \bar{Y}, \ldots, Y_J - \bar{Y})^T = \mathbf{H}\mathbf{Y}\,,$$

kjer je $\mathbf{H}$ centrirna matrika

$$\mathbf{H} = \mathbf{I} - \frac{1}{J}\,\mathbf{1}\mathbf{1}^T$$

in $\mathbf{Y} = (Y_1, \ldots, Y_J)^T$.

(ii) Neznano varianco σ^2 lahko ocenimo z

$$s_p^2 = \hat{\sigma}^2 = \frac{1}{I(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - Y_{i.})^2\,.$$

Pokažite, da ta ocena nepristranska.

(iii) Pokažite, da je

$$\frac{I(J-1)}{\sigma^2} s_p^2 \sim \chi^2(I(J-1))\,.$$

Pri tem uporabite Cochranov izrek, pri čemer upoštevajte, da je centrirna matrika $\mathbf{H}$ simetrična, idempotentna z rangom $J-1$. Prepričajte se, da trditev drži ne glede na to, ali H_0 velja ali ne, le predpostavka, da so populacijske variance enake v vseh skupinah, mora veljati.

(iv) Označite povprečje vseh Y_{ij} z

$$Y_{..} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} Y_{ij}\,.$$

Pokažite, da je izraz

$$s_b^2 = \frac{J}{I-1} \sum_{i=1}^{I} (Y_{i\cdot} - Y_{\cdot\cdot})^2$$

nepristranska cenilka σ^2, če H_0 drži. Dokažite še, da je

$$\frac{(I-1)}{\sigma^2} s_b^2 \sim \chi^2(I-1)\,.$$

Poleg tega utemeljite, da sta s_p^2 in s_b^2 neodvisni slučajni spremenljivki. Pri tem uporabite spet centrirno matriko $\mathbf{H}$ in točko (i).

(v) Idejo iz a., da primerjamo raztros vzorčnih povprečij z raztrosom znotraj skupin, lahko zdaj statistično "udejanimo". Pokažite, da je po definiciji

$$F = \frac{\frac{s_b^2}{\sigma^2}}{\frac{s_p^2}{\sigma^2}} \sim F_{I-1,I(J-1)}$$

c. Preberite razdelek 12.2.3 v učbeniku. Simulirajte porazdelitev testne statistike F in testne statistike K v dveh primerih za $I = 4$ in $J = 25$:

(i) V primeru, ko H_0 drži. Narišite graf empirične porazdelitve testnih statistik v obeh primerih.

(ii) V primeru, ko H_0 ne drži. V tem primeru primerjajte tudi moč F in K testa. Kruskal-Wallisov test ne potrebuje predpostavk o normalnosti. Na osnovi primerjave moči testov v primeru, ko porazdelitve Y_{ij} so normalne z enakimi variancami, komentirajte, kateri test bi bilo v praksi bolje uporabljati.