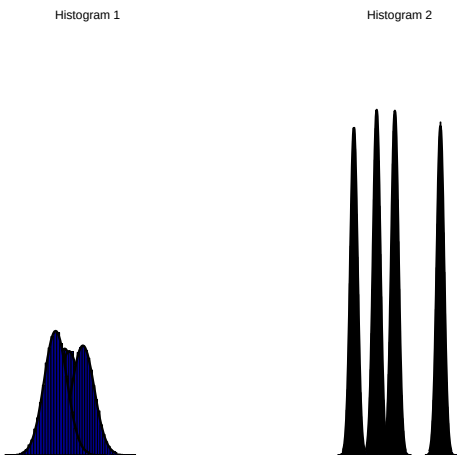
Analiza variance

Pri analizi variance v najpreprostejši obliki privzemamo, da imamo I različnih skupin. Populacijsko povprečje v teh skupinah označimo z μ_i, populacijsko variance pa z σ_i^2, $i = 1, 2, \ldots, I$. Iz vsake skupine izberemo enostavni slučajni vzorec **s ponavljanjem**. Privzemamo, da so vzorčne vrednosti slučajne spremenljivke Y_{ij}, ki so vse normalno porazdeljene, z $E(Y_{ij}) = \mu_i$ in $\mathrm{var}(Y_{ij}) = \sigma_i^2$. Preizkusiti želimo domnevo

$$H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_I \qquad \text{proti} \qquad H_1 \colon \text{vsi } \mu_i \text{ niso enaki.}$$

Neobhodna predpostavka pri analizi variance je, da so vse populacijske variance enake, torej $\sigma_1^2 = \cdots = \sigma_I^2$.

a. Na spodnji sliki sta dva možna histograma, če je $I = 4$. Oba kažeta vzorčne histograme za vsako skupino posebej.



Slika 1 Dva možna histograma za $I = 4$ skupine.

V katerem primeru je po vašem mnenju očitno, da so povprečja po posameznih skupinah različna? Zakaj?

b. Pri analizi variance "primerjamo" raztros vzorčnih povprečij, ki jih označimo z

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

z oceno raztrosa σ. Vemo, da v primeru, ko H_0 drži, vzorčna povprečja "variirajo" za $\sigma/\sqrt{J}$. Če so vzorčna povprečja "preveč raztresena" glede na primerjalno količino $\sigma/\sqrt{J}$, lahko H_0 zavrnemo. Potrebujemo pa nekaj teoretičnih osnov. Oglejte si kratek povzetek večrazsežne normalne porazdelitve na moji domači strani.

(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

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

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.