4.4 Testi võimsus

Mudelite võrdlemisel F-testi abil eeldasime, et keerukam mudel on kindlasti õige. Lihtsam kahest võrreldavast mudelist võis aga ei pruukinud olla õige. Nullhüpotees väitis, et ka lihtsam mudel on õige. Mis juhtub aga siis, kui lihtsam mudel pole õige? Kas ja millise tõenäosusega suudame F-testi abil tõestada, et lihtsam mudel pole õige? Alljärgnevalt arvutamegi, kui hea on F-testi võime tõestada alternatiivset hüpoteesi ehk milline on F-testi võimsus.

Dramatis personae:

- $\mathbf{M_0} : \mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$.
  Lihtsaim mudelitest. Nullhüpoteesi väide. Siin peatükis osutub valeks valikuks.
- $\mathbf{M_1} : \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$.
  Rikkam meie poolt kaalutav mudel. Õige mudel (kuid võib sisaldada mittevajalikke parameetreid, osad parameetervektori $\boldsymbol{\beta}_1$ elemendid võivad olla ka nullid).

Meenutame varem tõesatud teoreemi 3.1:

Olgu $\mathbf{y} \sim N(\boldsymbol{\mu}; \mathbf{I})$, ja olgu maatriks $\mathbf{A}$ ortogonaalne projektor (sümmeetriline ja idempotentne). Siis $\mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2(r, ncp)$, kus $r := rank(\mathbf{A}), ncp := \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$.

Ülalmainitud teoreemist järeldus (vaata näiteks valemit 3.1), et

$$\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{y}/\sigma^2 \sim \chi^2_{df=n-p_1}$$

ja juhul kui kehtib keerukam mudel, $\mathrm{E}(\mathbf{y}) = \mathbf{X}_1\boldsymbol{\beta}_1$:

$$\begin{aligned}
(\mathbf{y}/\sigma)^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})(\mathbf{y}/\sigma) &\sim \chi_{df=p_1-p_0;ncp=\mathrm{E}\mathbf{y}^T(\mathbf{P}_{\mathbf{X}_1}-\mathbf{P}_{\mathbf{X}_0})\mathrm{E}\mathbf{y}/\sigma^2} \\
\mathbf{y}^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}/\sigma^2 &\sim \chi_{df=p_1-p_0;ncp=\boldsymbol{\beta}_1^T\mathbf{X}_1^T(\mathbf{P}_{\mathbf{X}_1}-\mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1\boldsymbol{\beta}_1/\sigma^2}.
\end{aligned}$$

Juhul, kui õige on vaid keerukam mudel, siis antud avaldises mittetsentraalsuse parameeter ei muutu nulliks. Selgitame siinkohal veidi lähemalt mittetsentraalsuse parameetri arvutuseeskirja. Esmalt asendame kõik vaatlused oma andmestikus nende keskväärtustega. Seejärel leiame sellise vigadeta andmestiku pealt prognoosid samadele vaatlustele mõlema mudeli abil — nii lihtsama kui ka keerukama (õige) mudeli abil. Saadud prognooside erinevuste ruutude summa jagame jääkide dispersiooniga ja saamegi mittetsentraalsuse parameetri väärtuse.

Definitsioon 4.1 *Olgu*

$$X \sim \chi^2_{df_1;ncp}; \qquad Y \sim \chi^2_{df_2}$$

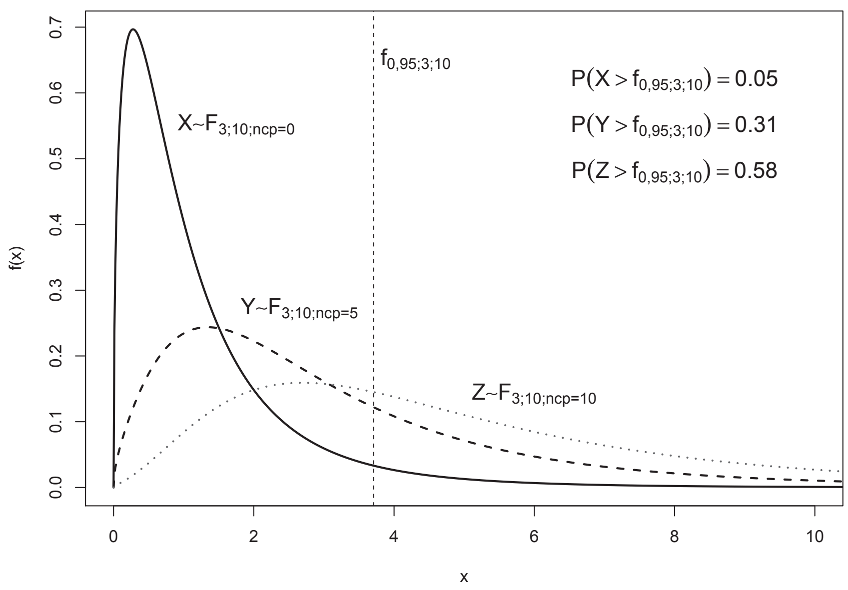
siis juhusliku suuruse

$$Z = \frac{X/df_1}{Y/df_2}$$

jaotuseks on mittetsentraalne F-jaotus vabadusastmete arvudega df_1, df_2 ja mittetsentraalsuse parameetriga ncp.

Tsentraalse ja mittetsentraalse F-jaotuse tihedusfunktsioonid on toodud ka joonisel 4.2.

Joonis 4.2: Tsentraalne ja mittetsentraalne F-jaotus



Seega tegelik F-statistiku jaotus juhul, kui õige on alternatiivne hüpotees (keerukam mudel) on

$$F \sim F_{df_1=p_1-p_0;df_2=n-p_1;ncp=\boldsymbol{\beta}_1^T\mathbf{X}_1^T(\mathbf{P}_{\mathbf{X}_1}-\mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1\boldsymbol{\beta}_1/\sigma^2}.$$

Kuna testi võimsus on tõenäosus kummutada nullhüpotees (ehk saada kriitilisest väärtusest suurem teststatistiku väärtus), siis võime testi võimsust leida valemiga

$$1 - F_{df_1=p_1-p_0;df_2=n-p_1;ncp=\boldsymbol{\beta}_1^T\mathbf{X}_1^T(\mathbf{P}_{\mathbf{X}_1}-\mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1\boldsymbol{\beta}_1/\sigma^2}(f_{0,95;p_1-p_0;n-p_1}),$$

kus $f_{0,95;p_1-p_0;n-p_1}$ tähistab tsentraalse F-jaotuse $df_1 = p_1-p_0 (= \mathrm{rank}(\mathbf{X}_1) - \mathrm{rank}(\mathbf{X}_0)); df_2 = n - p_1 (= n - \mathrm{rank}(\mathbf{X}_1))$ 0,95-kvantiili.

Näide 4.4 *Raha jätkub süvauuringute tegemiseks kümnele saarlasele, kümnele virumaalasele ja veerandsajale tartlasele (tartlaste puhul ei pea me maksma sõidu- ega ööbimiskulusi). Varasemate uuringute põhjal oletame, et uuritava tunnuse standardhälve on ligikaudu* $\sigma \approx 25$. *Uuritava tunnuse keskväärtused on ühe vastava ala tippspetsialisti arvates järgmised:* $\mu_{saarlased} = 100; \mu_{virulased} = 125; \mu_{tartlased} = 108$. *Milline on tõenäosus, et suudame F-testi abil tuvastada piirkondlike erinevuste olemasolu?*

Lahendus:
Lihtsama mudeli $Y_i = \mu + \varepsilon_i$ prognoos kõigile inimestele (kui inimeste keskväärtused oleksid täpselt teada) oleks üldkeskmine,

$$\hat{\mu} = (100 \cdot 10 + 125 \cdot 10 + 108 \cdot 25)/(10 + 10 + 25) = 110.$$

Keerukama mudeli prognoosiks oleks saarlastele saarlaste keskväärtus jne. Prognooside erienvuste ruutude summa tuleks

$$(100 - 110)^2 \cdot 10 + (125 - 110)^2 \cdot 10 + (108 - 110)^2 \cdot 25 = 3350$$

ja mittetsentraalsuse parameetriks saaksime $ncp = 3350/25^2 = 5{,}36$. Testi võimsuse leidmiseks peaksime veel tabelist või arvutist välja uurima F-jaotuse ($df_1 = 3 - 1$, $df_2 = 45 - 3$) kriitilise väärtuse ($f_{0,95;2;42} = 3{,}22$) ja võimegi leida, millise tõenäosusega kavandatava uuringu abil on võimalik tõestada alternatiivse hüpoteesi paikapidavust (testi võimsus):

$$1 - F_{df_1=2;df_2=42;ncp=5,36}(3{,}22) = 0{,}5027\ldots.$$

Kui kasutaksime kavandatud valimimahte, siis oleks regionaalsete erinevuste olemasolu tõestamise šansid ligikaudu 1:1.