Falskt öryggi stórra gagnasafna

Sigrún Helga Lund

prófessor í tölfræði við Háskóla Íslands

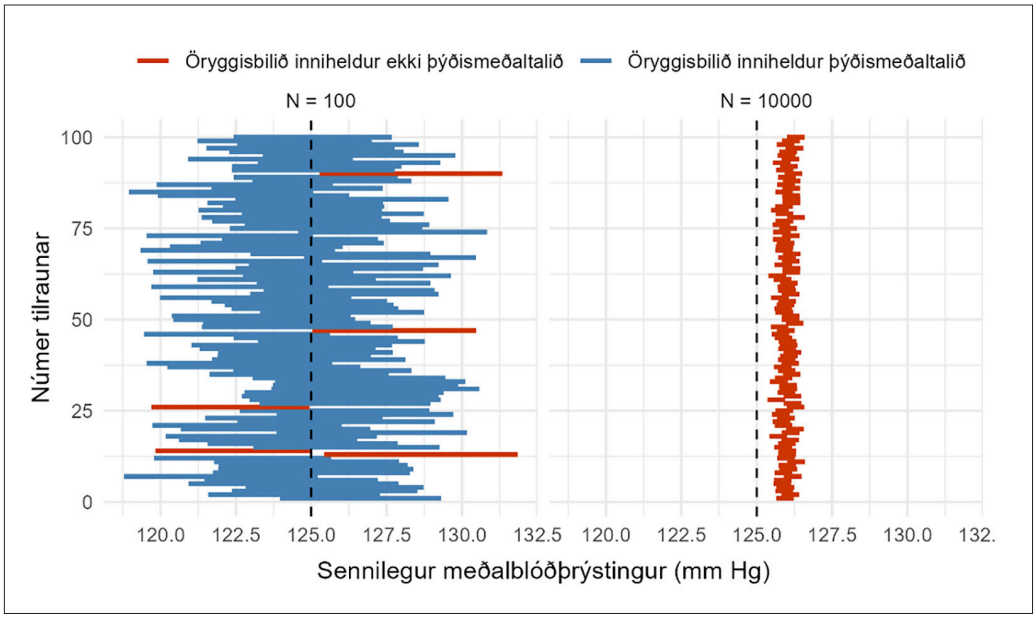
Nú til dags verður sífellt algengara að vísindamenn hafi aðgang að gríðarstórum gagnasöfnum, til dæmis úr rafrænum sjúkraskrám eða stórum lýðgrunduðum rannsóknum. Slík gagnasöfn veita mikið tölfræðilegt afl svo jafnvel er hægt er að greina smávægilegustu áhrif. Hins vegar eru þessi gögn sjaldnast fengin með handahófskenndu úrtaki úr þýðinu heldur hefur þeim verið safnað saman í gegnum klínískar mælingar, sjálfval þátttakenda eða önnur tilfallandi skilyrði. Þetta getur valdið úrtaksbjaga þar sem mælingarnar endurspegla ekki raunverulega dreifingu þýðisins. Úrtaksbjaginn getur í kjölfarið leitt til ýmissa bjagaðra mata. Í þessum pistli ætlum við að varpa ljósi á mögulega hættu slíkra bjaga með því að skoða öryggisbil.

Ímyndum okkur að við séum að meta meðaltal efri marka blóðþrýstings fullorðinna Íslendinga. Í fyrra tilvikinu tökum við handahófskennt úrtak af hundrað einstaklingum og mælum blóðþrýsting þeirra. Í seinna tilvikinu eru ekki gerðar nýjar mælingar heldur sóttar allar tiltækar blóðþrýstingsmælingar úr rafrænu sjúkraskrárkerfi stórrar heilsugæslustöðvar. Úrtakið er mun stærra og inniheldur tíu þúsund mælingar en gildi þessara mælinga eru ekki handahófsvalin úr þýði íslensku þjóðarinnar. Líklegt er að einstaklingar sem mæta á heilsugæslustöðina séu eldri eða veikari en almennt gerist, á meðan yngra og hraustara fólk er síður með aðgengilegar mælingar. Því er varlegt að gera ráð fyrir að meðaltal mælinga heilsugæslustöðvarinnar sé ívið hærra en hið raunverulega meðaltal í þýðinu.

Til að líkja eftir þessu skulum við gera ráð fyrir að efri mörk blóðþrýstings íslenska þýðisins séu normaldreifð með meðaltal 125 mm Hg og staðalfrávik 15 mm Hg. Við gerum sömuleiðis ráð fyrir að mælingarnar af heilsugæslustöðinni fylgi normaldreifingu en hækkum meðaltalið lítillega upp í 126 mm Hg og höldum staðalfrávikinu óbreyttu. Úr fyrri dreifingunni tökum við hundrað mælinga úrtak en úr þeirri seinni tíu þúsund mælinga úrtak. Fyrir bæði úrtökin reiknum við 95% öryggisbil. Þennan leik endurtökum við hundrað sinnum og má sjá niðurstöður allra hundrað tilraunanna á myndinni hér að neðan.

Vinstri hluti myndarinnar sýnir öryggisbil fyrir handahófsvöldu úrtökin af hundrað einstaklingum. Bilin eru breið og endurspegla meiri óvissu sem fylgir litlu úrtaki. Hins vegar eru þau rétt staðsett í kringum 125 mm Hg og innihalda oftast hið rétta þýðismeðaltal, að jafnaði í 95% tilvika. Hægri hluti myndarinnar sýnir öryggisbilin fyrir sjúkraskrárgögnin með tíu þúsund mælingum. Þessi bil eru mun þrengri en vegna úrtaksbjagans er þeim kerfisbundið hliðrað upp í kringum 126 mmHg. Ekkert þeirra inniheldur sanna þýðismeðaltalið.

Stór gagnasöfn hafa marga mikilvæga kosti. Þau gera okkur kleift að greina smávægileg áhrif sem hefðu annars farið framhjá okkur og bjóða uppá greiningu á ýmsum undirhópum með ásættanlegu tölfræðilegu afli. En þetta mikla afl felur í sér áhættu. Ef gagnasafnið er skekkt, endurspegla niðurstöðurnar það með enn nákvæmari hætti eftir því sem gögnin eru stærri. Þannig getum við endað með hárfínt mat á röngu gildi sem skapar falskt öryggi um skekktar niðurstöður. Meira er ekki alltaf betra.



Mynd 1. *Til vinstri, 95% öryggisbil frá slembivöldum hundrað mælinga úrtökum. Til hægri, 95% öryggisbil frá bjöguðum tíu þúsund mælinga úrtökum. Blár litur gefur til kynna að öryggisbilið innihaldi sanna þýðismeðaltalið, rauður litur gefur til kynna að öryggisbilið innihaldi ekki sanna þýðismeðaltalið.*