Predgovor

Na statističnih idejah in metodah temelji tako rekoč vsak vidik sodobnega življenja. Včasih je vloga statistike očitna, pogosto pa so statistične ideje in orodja skriti v ozadju. V vsakem primeru je zaradi vseprisotnosti statističnih idej očitno izredno koristno, da jih vsaj malo razumemo. Cilj te knjige je zagotoviti tako razumevanje.

Statistiki težavo povzroča nesrečno, a v temeljih zmotno prepričanje, ki ljudi zavaja o njeni resnični naravi. To zmotno mnenje je, da statistika zahteva obsežne suhoparne računske operacije in da je posledično dolgočasna in zaprašena disciplina, povsem brez domišljije, ustvarjalnosti ali razburljivosti. Vendar pa je to popolnoma napačna podoba sodobne discipline. Ta temelji na pojmovanju, ki sega več kot pol stoletja nazaj. Pri tem predvsem popolnoma prezre dejstvo, da je računalnik disciplino predrugačil in jo spremenil iz take, ki se je vrtela okoli računstva, v tako, ki v iskanju razumevanja in razjasnitve temelji na uporabi naprednih programskih orodij za obravnavanje podatkov. Prav to je bistvo sodobne discipline: uporaba orodij, ki pomagajo pri razumevanju in nudijo načine za osvetlitev, poti do razumevanja, sredstva za nadzor in vodenje ter sisteme za pomoč pri sprejemanju odločitev. Vse to in še veliko več so vidiki sodobne discipline.

Cilj te knjige je bralcu omogočiti vsaj osnovno razumevanje te sodobne discipline. Seveda je jasno, da v tako kratki knjigi, kot je ta, podrobnosti ne morem zajeti. Namesto podrobnega pregleda sem se odločil za pogled na celotno disciplino z višine, s ptičje perspektive, trudeč se prenesti naravo statistične filozofije, idej, orodij in metod. Upam, da bo knjiga bralcu

vsaj malo pojasnila, kako sodobna disciplina deluje, kako pomembna je in, seveda, zakaj je tako pomembna.

Prvo poglavje predstavi nekaj osnovnih definicij statistike in s pomočjo primerov pokaže del njene moči, pomembnosti in, da!, vznemirljivosti. Drugo poglavje predstavi nekatere od najbolj osnovnih statističnih idej, na katere je bralec povsem verjetno že naletel in ki se tičejo osnovnega povzemanja podatkov. Tretje poglavje nas opozori, da je veljavnost kateregakoli zaključka, ki ga povlečemo iz statističnih izračunov, bistveno odvisna od kakovosti neobdelanih podatkov, opiše pa tudi strategije za učinkovito zbiranje podatkov. Če so en temelj, na katerem stoji statistika, podatki, je drugi verjetnost; četrto poglavje tako opiše osnovne pojme verjetnosti. Z opisom, kako iz podatkov vleči sklepe in zaključke, v petem poglavju statistiko postavim na oba temelja. Šesto poglavje predstavlja bliskovit pregled nekaterih pomembnih statističnih metod in prikaže, kako so te del mreže medsebojno povezanih idej in metod za razbiranje pomena iz podatkov. Nazadnje sedmo poglavje predstavi nekaj načinov, kako je računalnik vplival na disciplino.

Za kritični pretres osnutkov te knjige bi se rad zahvalil Emily Kenway, Shelley Channon, Martinu Crowderju in anonimnemu bralcu. Njihovi komentarji so jo bistveno izboljšali in pomagali odpraviti nejasnosti v razlagah. Če je mestoma kakšna vendarle ostala, sem za to kriv popolnoma sam.

David J. Hand
Imperial College, London

1. poglavje
Obdani s statistiko

> Tistim, ki pravijo, da »obstajajo laži, preklete laži in statistika«, pogosto citiram Fredericka Mostellerja, ki je rekel, da »je s pomočjo statistike lahko lagati, še lažje pa brez nje«.

Sodobna statistika

Rad bi začel s trditvijo, ki se bo mnogim bralcem zdela presenetljiva: *statistika je ena najvznemirljivejših disciplin*. Cilj te knjige je dokazati resničnost te trditve in pokazati, zakaj je resnična. Upam, da bom uspešno ovrgel nekatere od starih zmotnih predstav o naravi statistike in pokazal, kakšna je sodobna disciplina, obenem pa predstavil tako nekaj njene vrtoglave moči kot njeno vseprisotnost.

Še posebej pa bi v tem poglavju rad predstavil dvoje stvari. Prva je okus po revoluciji zadnjih nekaj desetletjih. Razložil bom, kako se je statistika preoblikovala iz suhoparne viktorijanske discipline, ki se ukvarja z ročnim obdelovanjem številskih stolpcev, v izredno prefinjeno sodobno tehnologijo, ki vključuje uporabo najnaprednejših programskih orodij. Ponazoril bom, kako sodobni statistiki ta orodja uporabljajo za obravnavanje podatkov v iskanju struktur in vzorcev in kako to tehnologijo uporabljajo za odstiranje plasti zameglitev in nejasnosti, pri tem pa razkrivajo resnico za njimi. Sodobna statistika nam kot teleskopi, mikroskopi, rentgenski žarki in medicinsko slikanje omogoča videti stvari, ki so prostim očem nevidne. Omogoča nam videti skozi koprene in zmešnjavo sveta okoli nas, da dojamemo prikrito stvarnost.

To je torej prva stvar, ki bi jo rad razložil v tem poglavju: sáma moč in vznemirljivost sodobne discipline, od kod je prišla in kaj je sposobna narediti. Druga stvar, ki bi jo rad razložil, pa je vseprisotnost statistike. V sodobnem življenju ni niti enega vidika, ki se ga ne bi dotikala. Sodobna medicina temelji na statistiki: randomiziran nadzorovan poskus je bil denimo opisan kot »eno najpreprostejših, najmočnejših in najbolj revolucionarnih raziskovalnih orodij«. Razumevanje načinov širjenja epidemij prepreči, da bi te zdesetkale človeštvo. Učinkovita vlada je odvisna od pazljive statistične analize podatkov, ki opisujejo gospodarstvo in družbo: morda je to argument za zahtevo, naj vsi tisti, ki so v vladi, obvezno opravijo tečaj iz statistike. Kmetje, živilski tehnologi in lastniki supermarketov pri odločanju o tem, kaj posaditi, kako to obdelati in kako obdelano zapakirati in razposlati, vsi implicitno uporabljajo statistiko. Hidrologi se z analiziranjem meteoroloških podatkov odločajo, kako visoko zgraditi protipoplavne pregrade. Inženirji, ki razvijajo računalniške sisteme, z uporabo statistike zanesljivosti zagotovijo, da se sistemi ne sesuvajo prepogosto. Sistemi za nadzor zračnega prometa so zasnovani na podlagi kompleksnih statističnih modelov, ki delujejo v realnem času. Čeprav tega morda ne prepoznate, so statistične ideje in orodja skriti v praktično vsakem vidiku sodobnega življenja.

Nekaj definicij

Dobra delovna definicija statistike bi lahko bila, da je statistika *tehnologija za razbiranje pomena iz podatkov*. A nobena definicija ni popolna. Ta zlasti ne upošteva slučajnosti in verjetnosti, ki sta stebra za množico načinov uporabe statistike. Tako bi se lahko druga delovna definicija glasila, da je statistika *tehnologija ravnanja z negotovostjo*. Natančnejše definicije bi morda dale večji poudarek različnim vlogam statistike. Tako bi lahko rekli, da je statistika ključna disciplina za *predvidevanje*

prihodnosti ali za *sklepanje o neznanem* ali za *prikladno povzemanje podatkov*. Vse te definicije skupaj v grobem pokrijejo bistvo discipline, čeprav bodo različne uporabe statistike določale zelo različne pojavne oblike. Sprejemanje odločitev, napovedovanje, nadzor v realnem času, odkrivanje goljufij, popis štetja in analiza genetskih zaporedij – vse to so primeri uporabe statistike, ki pa lahko zahtevajo zelo različne metode in orodja. Pri naštetih definicijah bodite pozorni na to, da sem namenoma izbral besedo »tehnologija« in ne znanost. Tehnologija je uporaba znanosti in njenih odkritij in prav to počne tudi statistika: gre za uporabo našega razumevanja, kako iz podatkov izluščiti informacije, in za uporabo našega razumevanja negotovosti. Kljub temu se o statistiki včasih govori kot o znanosti. Pravzaprav se tako imenuje celo ena najzanimivejših statističnih revij: *Statistical Science* oziroma *Statistična znanost*.

Doslej sem se v tej knjigi, sploh v predhodnem odstavku, skliceval na *disciplino statistike*, vendar pa ima beseda statistika še en pomen: pomeni tudi zbir statističnih podatkov. Statistični podatek je numerično dejstvo ali povzetek. Lahko je denimo povzetek podatkov, ki opisujejo neko populacijo: morda njeno velikost, rodnost ali stopnjo kriminala. Tako je na neki način to tudi knjiga o posameznih numeričnih dejstvih. A v resnici je veliko več kot to. Je knjiga o tem, kako zbirati numerična dejstva, kako jih obravnavati, analizirati in iz njih sklepati. Je knjiga o sami tehnologiji. Bojim se, da bo bralec, ki upa, da bo v tej knjigi našel tabele s števili (na primer »športno statistiko«), razočaran. Nagrajen pa bo bralec, ki želi razumeti, kako podjetja sprejemajo odločitve, kako astronomi odkrivajo nove vrste zvezd, kako raziskovalci na področju medicine odkrivajo gene, ki so povezani z določenimi boleznimi, kako se banke odločajo nekomu dati kreditno kartico ali ne, kako zavarovalnice postavljajo cene premij, kako lahko ustvarimo filtre za vsiljeno pošto, ki preprečijo, da bi opolzke reklame prišle v naš e-poštni predal, in tako dalje in tako dalje.