1. JOHDATTELEVIA ESIMERKKEJÄ

AS = Arjas - Sirén moniste

1. Pallot kulhossa; vrt. [AS, jakso 1.1]

Kulhossa N palloa: K keltaista ja $N-K$ valkoista. Tark. satunnaiskoetta

(*) { nostetaan umpimähkään ja palauttaen (sekoittaen välillä hyvin sekoittaen) n palloa

Pal. tn-laskennasta mieleen, miten tätä voidaan mallintaa tn-laskennan avulla: Määr. sm:t

$$X_i = \begin{cases} 1, & \text{jos } i\text{:s pallo keltainen} \\ 0, & \text{jos } i\text{:s pallo valkoinen} \end{cases} \qquad (i = 1, \ldots, n)$$

$\Rightarrow$ $X_1, \ldots, X_n$ $\perp\!\!\!\perp$ (riippumattomia) ja kullakin X_i:llä on pistetodennäköisyydet (lyh. ptn)

$$\begin{cases} P(X_i = 1) = K/N = \theta \\ P(X_i = 0) = 1 - K/N = 1 - \theta \end{cases} \qquad (1)$$

jossa $\theta = K/N$ keltaisten suht. osuus kulhossa.

Merk. $\underset{\sim}{X} = (X_1, \ldots, X_n)$ ("n-ulott. satunnaisvektori")

Kun $\underset{\sim}{x} = (x_1, \ldots, x_n)$ jossa kukin $x_i = 0$ tai 1, saadaan riippumattomuuden nojalla (yhteys) ptn:t

$$\boxed{P(\underset{\sim}{X} = \underset{\sim}{x}) = P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) \overset{(1)}{=} \theta^t (1-\theta)^{n-t}} \qquad (2)$$

jossa

$$\begin{cases} t = x_1 + \cdots + x_n = \text{keltaisten lkm "otoksessa"} \\ n - t = \text{valkoisten lkm "otoksessa"} \end{cases}$$

Jos kiinnitetään huomiota vain keltaisten lkm:ään otoksessa eli sm:aan $T = X_1 + \cdots + X_n$, tiedetään tn-laskennasta, että $T \sim \mathrm{Bin}(n, \theta)$ eli T:llä on ptn:t

$$\boxed{P(T = t) = \binom{n}{t} \theta^t (1-\theta)^{n-t}, \quad t = 0, 1, \ldots, n} \qquad (3)$$

Sitspä: Edellyttäen, että tunnemme keltaisten suht. osuuden θ kulhossa, kaavat (2) ja (3) kertovat täydellisesti kokeen (x) eri tulosvaihtoehtojen tn:t (riippuen siitä haluammeko pitää kokeen "tuloksena" täydellistä tulospöytäkirjaa $\underset{\sim}{x} = (x_1, \ldots, x_n)$ vai arvoastaan keltaisten lkm:ää t).

Tilastollinen päättely tutkii käänteistä ongelmaa: Oletetaan, että θ on tuntematon ja suoritetaan koe (x). Havaitaan

- joko $\underset{\sim}{x} = (x_1, \ldots, x_n)$ eli nostettujen pallojen värit järjestyksessä
- tai vain keltaisten lkm $t = x_1 + \cdots + x_n$.

Mitä voidaan päätellä θ:sta tämän perusteella ?

Esim. Suoritetaan $n = 100$ nostoa ja havaitaan $t = 35$ keltaista. Mitä johtopäätöksiä θ:sta ?

(2.) Kliininen koe. Tautiin on kehitetty lupaava lääke, jota halutaan testata. Oletetaan:

- ilman lääkettä taudista paranee tn:llä θ_0
- lääkkeellä hoidettuna siitä paranee tn:llä θ_1

Jaetaan 100 potilasta satunnaisesti kahteen 50 hengen ryhmään. Toiselle annetaan placeboa, toiselle lääkettä.

Seurantajakson kuluttua havaitaan:

- lumeboryhmässä parani x_0 potilasta
- lääkeryhmässä parani x_1 potilasta

Mitä voidaan <u>päätellä</u> luvuista θ_0 ja θ_1?
Erityisesti: voidaanko päätellä, että $\theta_1 > \theta_0$ (kuten tutkija luultavasti toivoo!)?

(3.) <u>Metron odotusajat</u>. Oletetaan, että metro kulkee säännöllisesti θ minuutin väliajoin. Umpimähkään laiturille saapuvan henkilön odotusajan voidaan siis ajatella olevan $\mathrm{Tas}(0, \theta)$ - jakautunut.

Maalta muuttanut opiskelija ei tunne θ:aa, mutta joka kerta laiturille saavuttuaan hän mittaa odotusaikansa, saaden siten mittaustulokset $t_1, \ldots, t_n$ (minuuttia). Mitä hän voi <u>päätellä</u> θ:sta?

<u>Esim.</u> Kolmen kerran tulokset ovat $t_1 = 3.4$, $t_2 = 1.1$ ja $t_3 = 4.5$.

- Täysin varmasti voi päätellä, että $\theta \geq 4.5$.
- Kuinka luotettavasti voisi päätellä, että $\theta \leq 10$?

(4.) <u>Appelsiinikauppa</u>. Hedelmäkauppias väittää, että hänen myymiensä appelsiinien keskipaino μ on ainakin 250 g. Asiakas ostaa umpimähkään 10 appelsiinia ja havaitsee kotona, että niiden keskipaino on 236 g.

Onko tämän perusteella syytä asettaa kyseenalaiseksi kauppiaan väite "$\mu \geq 250$" vai kävikö vain "huono tuuri"?

Voiko asiaan edes ottaa kantaa ilman lisätietoja/-oletuksia?

Pohdittavaksi yo. "tutkimusasetelmiin" liittyen:

* Mitkä olivat kokeellisia, mitkä puhtaasti havainto-tutkimuksia? Mitkä olivat otantatutkimuksia?

* Olivatko havaintoja vastaavat muuttujat diskreettejä vai jatkuvia?

* Miten todennäköisyyden käsite on tulkittava? Klassisesti (symmetriset alkeistapaukset)? Frekventistisesti? Jotenkin muuten?

2. TILASTOLLISEN PÄÄTTELYN YLEINEN ASETELMA JA TAVOITTEET

Tarkasteltavana aineisto $\underset{\sim}{x} = (x_1, \ldots, x_n)$, joka koostuu havainnoista $x_1, \ldots, x_n$. Niihin liittyy satunnaisuutta - johtuen esim. satunnaisotannasta tai tutkittavan ilmiön itsensä satunnaisesta luonteesta

$\Rightarrow$ ajattelemme että ne ovat eräiden satunnaismuuttujien

$$X_1, \ldots, X_n \quad ; \quad \text{vektorina} \quad \underset{\sim}{X} = (X_1, \ldots, X_n)$$

toteutuneita arvoja eli realisaatioita.

Näiden sm:ien (yhteis)jakauma $f(\underset{\sim}{x}) = f_{\underset{\sim}{X}}(\underset{\sim}{x})$ (käytännössä: yhteispistetn- tai yhteistiheysfunktio) on (ainakin osittain) tuntematon ja tavoitteena on tehdä siitä päätelmiä aineiston $\underset{\sim}{x}$ perusteella.

Useimmiten (ja erit. tällä kurssilla) oletetaan, että jakauma on tunnettu jotakin parametria θ vaille, ts. $\underset{\sim}{X}$:n yptnf/ytf on muotoa

$$f(\underset{\sim}{x}; \theta) = f_{\underset{\sim}{X}}(\underset{\sim}{x}; \theta),$$

jossa θ on tuntematon.

Määr. $f(\underset{\sim}{x}; \theta)$ on (parametrinen) tilastollinen malli.

Tässä parametri θ voi olla reaaliarvoinen (1-ulott.) tai koostua useammasta komponentista: $\theta = (\theta_1, \ldots, \theta_k)$

Yleinen tavoite: tehdä johtopäätöksiä θ:sta havaitun aineiston $\underset{\sim}{x}$ perusteella.

Erityisluonteisempia:

(1.) Piste-estimointi: määritettävä piste (luku tai vektori) kaikkien mahdollisten parametriarvojen joukosta, joka on "hyvä arvaus" θ:n todelliselle "oikealle" arvolle.

(2.) Väliestimointi: rajattava sellainen parametriarvojen joukko (1-ulott. tapauksessa yleensä väli), joka suurella varmuudella sisältää "oikean" θ:n arvon.

(3.) Hypoteesintestaus: onko aineisto sopusoinnussa annetun hypoteesin "$\theta \in H$" kanssa vai ei?

Muuta til. päättelyn piiriin kuuluvaa:

(4.) Ennustaminen: Aineiston $(x_1, \ldots, x_n)$ perusteella ennustettava samasta satunnaisilmiöstä/kokeesta tulevaa uutta/seuraavaa havaintoa x_{n+1}.

(5.) Mallinarviointi (diagnostiikka): Onko malli $f(\underset{\sim}{x}; \theta)$ sopiva ja oikea selittämään saadut havainnot? Toteutuvatko mallin oletukset?

Huom. Johainen malli on aina todellisuuden yksinkertaistus ja approksimaatio!

Tällä kurssilla lähinnä aiheet 1. – 3.
Käytännön sovelluksissa tehtävä 5. hyvin tärkeä !

Esim. "Pallot kulhossa" -esimerkissä edellä tilasl. malli sm:lle T on

$$T \sim \mathrm{Bin}(n, \theta)$$

$(\Leftrightarrow)$

$$f(t;\theta) = P(T=t) = \binom{n}{t} \theta^t (1-\theta)^{n-t}, \quad t = 0, 1, \ldots, n$$

Parametri on θ, $0 \le \theta \le 1$ (keltaisten suht. osuus)
(n = nostojen lkm ajatellaan tunnetuksi luvuksi)

Päättelyn kaksi tärkeää paradigmaa/koulukuntaa:

Frekventistinen päättely (kurssin alkuosassa)

- Aineisto $\underset{\sim}{x}$ on satunnaisvektorin $\underset{\sim}{X}$ toteutunut arvo.
- Satunnaisuus viittaa "toistetun aineistonkeruun" ideaan: frekventistinen tn:n tulkinta
- θ on kiinteä mutta tuntematon luku tai piste
- θ:lla ei ole todennäköisyysjakaumaa !

Bayesiläinen päättely (kurssin loppupuolella)

- Myös parametri tulkitaan satunnaismuuttujaksi
- Todennäköisyys kuvaa siihen liittyvää epävarmuutta, subjektrivinen tn:n tulkinta
- Tyylikäs tapa yhdistää parametria koskeva ennakkotieto ja aineiston antama lisäinformaatio.
- Perustuu Bayesin kaavan käyttöön