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

Esim. "Pallot kulhossa" -esimerkissä edellä tilast. 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 !

Bayesläinen päättely (kurssin loppupuolella)

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

19.3.

3. TILASTOLLISEN MALLIN MUODOSTAMISESTA

Yleistä:

- Ei helppo tehtävä (aivan perusesimerkkejä lukuunottamatta)!
- Edellyttää ilmiöön liittyvän taustateorian tuntemusta.
- Usein jatkuva ja iteratiivinen prosessi (vrt. 5 sivulla 5)
- Meillä mallit joko helposti muodostettavia tai "valmiiksi annettuja".

Kursta: Malli $f(\underset{\sim}{x};\theta)$ (tai selvyyden vuoksi $f_{\underset{\sim}{X}}(\underset{\sim}{x};\theta)$) on

* diskreetissä tapauksessa (yhteis) pistetnf

$$f(\underset{\sim}{x};\theta) = P_\theta(\underset{\sim}{X} = \underset{\sim}{x}) = P_\theta(X_1 = x_1, \ldots, X_n = x_n)$$

* jatkuvassa tapauksessa (yhteis) tiheysfunktio, ts.

$$P_\theta(\underset{\sim}{X} \in A) = \int_A f(\underset{\sim}{x};\theta)\, d\underset{\sim}{x}, \quad \text{kun} \quad A \subset \mathbb{R}^n,$$

(n-uloitt. integraali) $\qquad \underset{\sim}{x} = (x_1, \ldots, x_n)$

(Tarkempi käsittely aineopintojen tn-laskennan ja päättelyn kursseilla.)

Erikoistapaus: Riippumattomat samoin jakautuneet havainnot:

$$\begin{cases} X_1, \ldots, X_n \ \perp\!\!\!\perp \\ X_i \text{ :llä ptnf/tf } g(\cdot\,;\theta) \quad \text{(sama jokaisella } i\text{)} \end{cases}$$

Tällöin tn-laskennan perusteella

$$f(\underset{\sim}{x};\theta) = g(x_1;\theta) \cdots g(x_n;\theta)$$

$$= \prod_{i=1}^{n} g(x_i;\theta), \qquad \underset{\sim}{x} = (x_1, \ldots, x_n)$$

Huom. Riippumattomuusoletus ei aina toteudu! Esimerkkinä aikasarja-tyyppiset mallit, joissa havainnot $X_1, \ldots, X_n$ ovat saman muuttujan arvoja peräkkäisinä ajanhetkinä. Ajattele esim.

X_i = lämpötila Kaisaniemessä vuoden i:ntenä päivänä klo 12.00

Esim. Riippumaton otos normaalijakaumasta $N(\mu, \sigma^2)$:

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2) \quad \perp\!\!\!\perp$$

"kaikkien tilast. mallien äiti"

Parametri 2-ulotteinen (μ, σ^2), $\mu \in \mathbb{R}$, $\sigma^2 > 0$.
(Joskus σ^2 tunnettu, jolloin parametrina vain μ.)

Pal. mieleen tn-laskennasta: $N(\mu, \sigma^2)$-jakauman tf

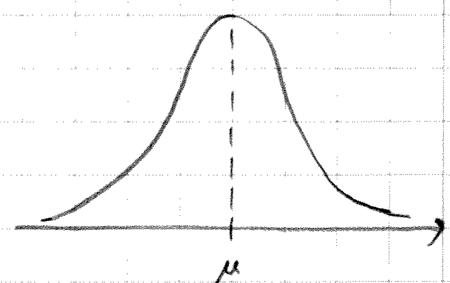
$$g(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$$

Siten tilast. mallin lauseke (yhteistf) on

$$f(\underset{\sim}{x}; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x_i-\mu)^2 \right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2 \right\}$$

Palaamme tähän myöhemmin.



Muista: Jos $X \sim N(\mu, \sigma^2)$, niin
$\mu = E(X)$ odotusarvo ja
$\sigma^2 = \mathrm{Var}(X) = D^2(X)$ varianssi

Muita paljon käytettyjä jakaumia: (tn-laskennan kurssi!)

* $X \sim \mathrm{Poisson}(\lambda)$ Poisson-jakauma parametrina $\lambda > 0$, pistetnf

$$g(x; \lambda) = P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

* $X \sim Exp(\lambda)$ <u>eksponenttijakauma</u> parametrina $\lambda > 0$, tiheysf

$$g(x;\lambda) = \lambda e^{-\lambda x}, \quad x > 0$$

* <u>tasajakauma</u> välillä $[\alpha, \beta]$, $Tas(\alpha, \beta)$

4. USKOTTAVUUSFUNKTIO JA SUURIMMAN USKOTTAVUUDEN ESTIMAATTI

vrt. Arjas - Sirén jakso 1.2.

<u>Motivaatio:</u> Tark. mallia $T \sim Bin(n, \theta)$, parametri $0 \leq \theta \leq 1$.

<u>esim.</u> "nostetaan n palloa kulhosta palauttaen", $\theta = \left\{\begin{array}{l}\text{keltaisten}\\ \text{osuus}\end{array}\right.$

T = keltaisten lkm otoksessa

tai <u>yleisemmin</u>: n-kertainen riippumaton <u>toistokoe</u>, jossa kunkin toiston "onnistumistn" $= \theta$

T = "onnistumisten" lkm

(A) Olet. että $\boxed{n = 10}$ ja havaittu $\boxed{T = 6}$

Ko. havainnon todennäköisyys on (ks. s. 6)

$$P(T=6) = f(6;\theta) = \binom{10}{6} \theta^6 (1-\theta)^4$$

Tark. tätä θ:n funktiona: (ks. kuvaaja sivulla 11)

$$L(\theta) = 210 \cdot \theta^6 (1-\theta)^4, \quad 0 \leq \theta \leq 1$$

Siis: $L(\theta) = P_\theta(T = 6)$ = tn saada havainto "$T = 6$" silloin kun parametrilla arvo θ

<u>suurimmillaan</u> L on pist. $\theta = 0.6$: $L(0.6) \approx 0.25$

sanomme: $\theta = 0.6$ on (havainnon $T = 6$ valossa) <u>uskottavin</u> parametrinarvo, koska se maksimoi ko. havainnon saamisen tn:n!

toisaalta esim. $L(0.4) \approx 0.1$, joten mikäli parametrilla on arvo $\theta = 0.4$, on havainnon "$T = 6$" saaminen ~ 2.5 kertaa epätodennäköisempää kuin siinä tapauksessa että $\theta = 0.6$

sanomme: $\theta = 0.6$ on ~ 2.5 kertaa uskottavampi parametrinarvo kuin $\theta = 0.4$

jne...

Huom. L:n kuvaaja on varsin "laakea": "melko uskottavia" parametrinarvoja on paljon
$\Rightarrow$ tarkkoja ja luotettavia päätelmiä θ:sta ei mahdollista tehdä! Ei ihme, sillä otoskoko $n = 10$ on hyvin pieni!

(B) olet. $\boxed{n = 300}$ ja havaittu $\boxed{T = 159}$ (luennolla tehty koe!)

Menetellään kuten edellä: saadun havainnon tn on nyt

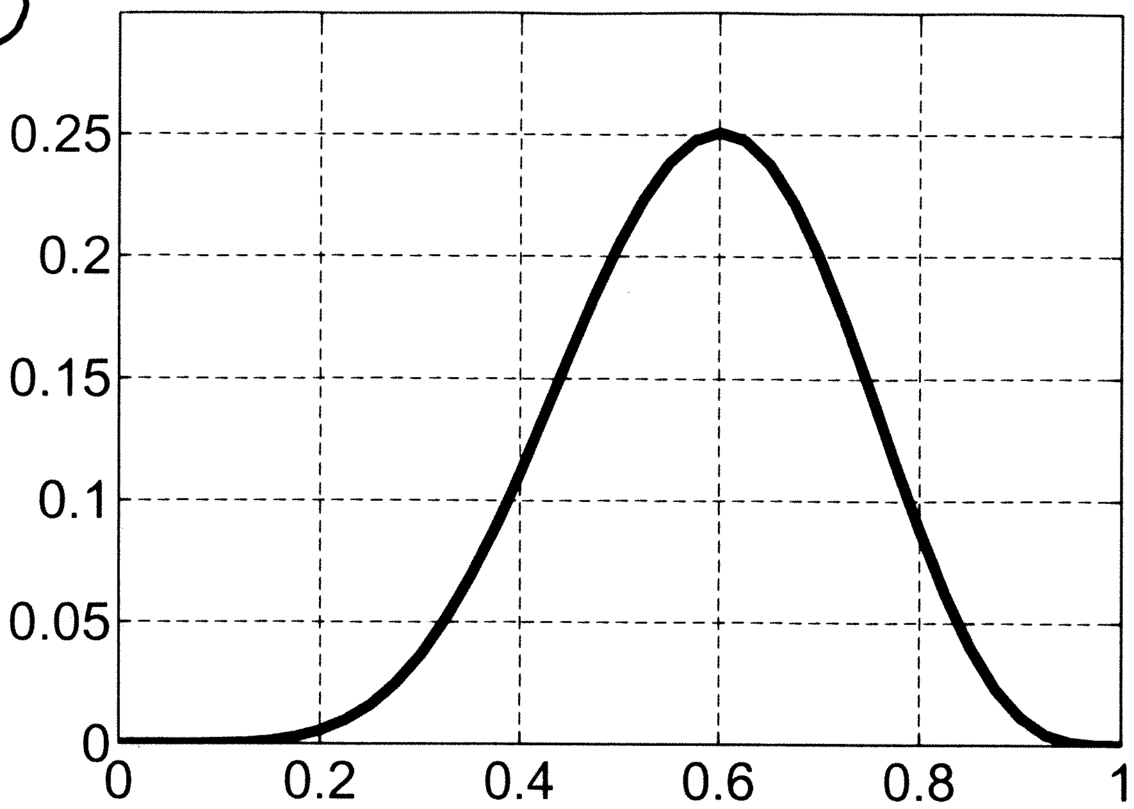
$$L(\theta) = P_\theta(T = 159) = \binom{300}{159} \theta^{159} (1-\theta)^{141}$$

(ks. kuvaaja sivulla 11)

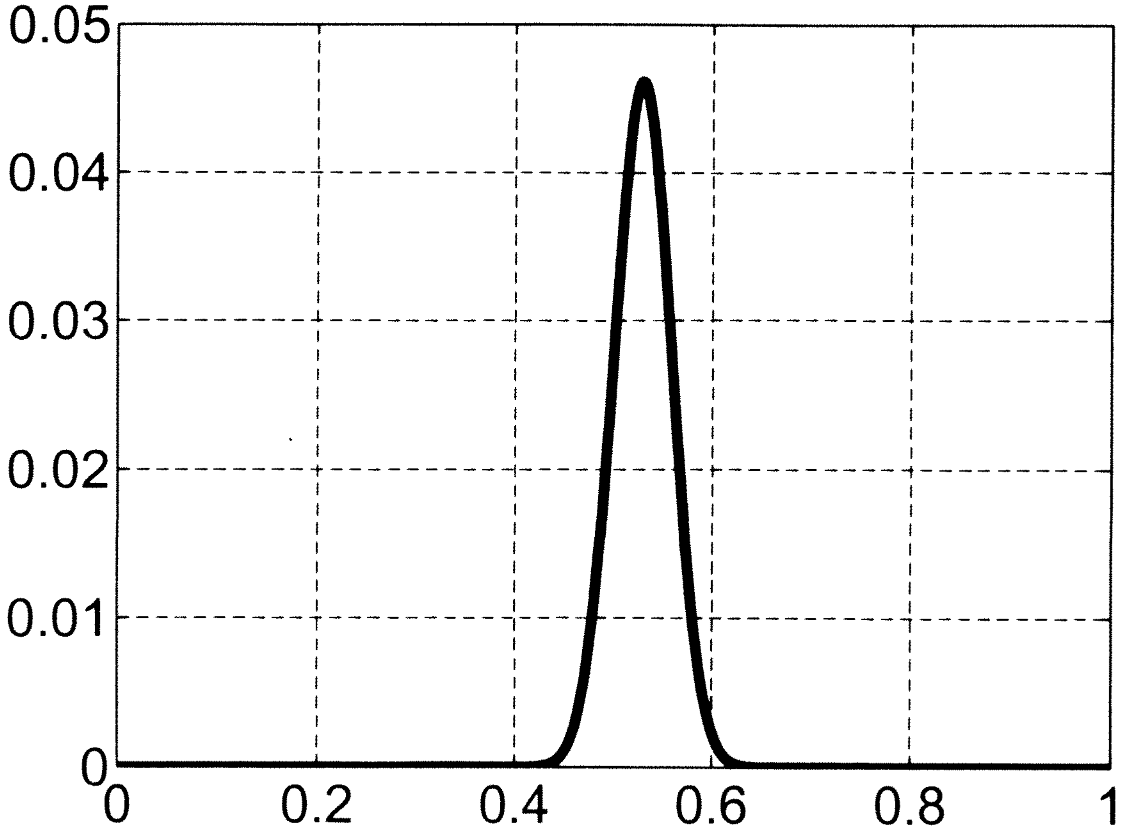
Nyt L:n globaali maksimikohta on $\theta = \frac{159}{300} = 0.53$, siis tämä on uskottavin parametrinarvo.
[Tarkistamme tämän kohta differentiaalilaskennan avulla!]

Huom. L:n kuvaaja "huipukkaampi" kuin (A)-tilanteessa ts. "uskottavat" parametrinarvot ovat melko kapealla välillä 0.53:n ympärillä $\Rightarrow$ θ:sta voi tehdä tarkempia päätelmiä kuin (A):ssa (suuremmasta otoskoosta johtuen).

A


0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1

B


0.05
0.04
0.03
0.02
0.01
0
0
0.2
0.4
0.6
0.8
1

Määritelmiä. Olkoon $f(\underset{\sim}{x};\theta)$ tilastollinen malli, jonka parametriavaruus (so. parametrin θ kaikkien mahdollisten arvojen joukko) on Ω.

* Aineistoon $\underset{\sim}{x} = (x_1, \dots, x_n)$ liittyvä uskottavuusfunktio on

$$L(\theta) = L(\theta; \underset{\sim}{x}) = f(\underset{\sim}{x}; \theta), \quad \theta \in \Omega$$

* Jos $\theta \in \Omega$ ja $\theta' \in \Omega$ siten että $L(\theta; \underset{\sim}{x}) > L(\theta'; \underset{\sim}{x})$, sanomme, että θ on (aineiston $\underset{\sim}{x}$ valossa) uskottavampi parametrinarvo kuin θ'.

* Sellainen piste $\hat{\theta} = \hat{\theta}(\underset{\sim}{x}) \in \Omega$, jossa L saavuttaa suurimman arvonsa, ts. jossa

$$L(\hat{\theta}; \underset{\sim}{x}) \geq L(\theta; \underset{\sim}{x}) \quad \forall\, \theta \in \Omega,$$

on parametrin θ suurimman uskottavuuden estimaatti (lyh. su-estimaatti).

Tulkinta: Su-estimaatti on sellainen parametrinarvo, jonka vallitessa "käsilläolevan" havainnon $\underset{\sim}{x}$ saamisen todennäköisyys (tai jatkuvan jak. tapauksessa "tn-tiheys") on suurin.

Huom. L ei ole tn-jakauma (θ ei ole sat. muuttuja)!

Esim. Palataan toistokoemalliin $T \sim Bin(n, \theta)$ (parametri θ) eli

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

Havaintoa t vastaava uskottavuusfunktio on

$$L(\theta) = L(\theta; t) = \binom{n}{t} \theta^t (1-\theta)^{n-t}, \quad 0 \leq \theta \leq 1.$$

Su-estimaatin laskemiseksi tutkitaan tämän logaritmia: