T-61.281 Luonnollisen kielen tilastollinen käsittely

Harjoitus 1, ti 27.1.2004, 8:15-10:00 – Palautellaan mieliin todennäköisyyslaskuja
Versio 1.0

1. Englannin kielessä voisi päteä seuraavanlaiset todennäköisyydet:

   $P($ sana=lyhenne | sana=kolmikirjaiminen $) = 0.8$

   $P($ sana=kolmikirjaiminen $) = 0.0003$

   Millä todennäköisyydellä satunnainen havaittu sana on kolmikirjaiminen lyhenne ?

2. Pikkunäppärä todennäköisyyspähkinä:
   Oope Rankka haluaa opettaa veljenpoikiaan neuvokkaiksi liikemiehiksi. Hän ehdottaa pojille peliä: Hän laittaisi kunkin pojan päähän satunnaisesti sinisen tai punaisen lippiksen ja laskisi kolmeen. Kolmen kohdalla kunkin pojan olisi oltava hiljaa tai arvattava ääneen oman lakkinsa väri. Jos kukaan ei arvaisi väärin ja yksikin arvaisi oikein, pojat saisivat euron arvoiset jätskit. Pelin aikana tietysti kaikenlainen merkinanto olisi kiellettyä, pojat näkisisivät vain toisten lakit, mutteivät omaansa.

   Veljenpojat Kupu, Rupu ja Pupu supattivat keskenään hetken ja suostuvat peliin. Voitonriemuisena Pupu vielä toteaa: " On meil ainaskin yli puolen mahollisuus voittaa jätskit !".

   Miten tähän hämmästyttävään tulokseen päästään, vai puhuuko Pupu vain lämpimikseen ? Kuinka paljon poikien kannattaisi maksaa saadakseen osallistua tähän peliin ?

3. Tarkastellaan lingvisti Å. Lindquistin kehittämää sanan perusmuotoistuskonetta. Kontekstin perusteella se osaa johtaa sanan *"siitä"* perusmuodoksi joko sanan *"se"* tai *"siittää"*. Laite osaa päätellä perusmuodosta *"se"* taivutetun sanan oikean perusmuodon todennäköisyydellä 0.95 ja väärä perusmuoto lipsahtaa todennäköisyydellä 0.05. Samoin käy perusmuodosta *"siittää"* taivutetuille sanoille. Koska perusmuoto *"se"* on paljon yleisempi, vain joka tuhannes *"siitä"* pitää perusmuotoistaa sanaksi *"siittää"*.

   Laite kertoo meille, että erään sanan *"siitä"* perusmuoto on *"siittää"*. Millä todennäköisyydellä laite on oikeassa ?

4. Kun lasketaan kielestä yksinkertaisia tilastoja, viitataan usein Zipfin lakiin. Sanat taulukoidaan niin, että yleisin laitetaan ensimmäiseksi ($r = 1$) ja muut järjestyksessä sen perään ($r = 2, 3, \ldots$). Kunkin sanan viereen kirjoitetaan kuinka monta kertaa se esiintyi tekstissä (f). Zipf väittää että

   $$f \propto \frac{1}{r}$$

Sanallisesti sanottuna siis f on verrannollinen $\frac{1}{r}$:ään tai $f * r$=vakio.

Päteekö Zipfin laki satunnaisesti generoidulle kielelle, jossa on 30 kirjainta, joista yksi on sanaväli ?

5. a) Heitetään 101-sivuista noppaa, jonka sivuilla on luvut 0 – 100. Laske saadun silmäluvun odotusarvo ja varianssi. Hahmottele todennäköisyyden p(X) kuvaaja jossa X on heiton silmäluku.

   b) Heitetään kahta 101 sivuista noppaa ja jaetaan silmälukujen summa kahdella. Laske tuloksen odotusarvo ja varianssi. Hahmottele todennäköisyyden p(X) kuvaaja jossa X on heiton silmälukujen summa jaettuna kahdella.

   c) Ja vielä heitetään kymmentä noppaa, jaetaan tulos kymmennellä. Hahmottele kuvaaja.

   d) Etsitään käsiimme kaikki maailman nopat ($n \to \infty$). Millainen jakauma meillä nyt mahtaa olla ? Hahmottele kuvaaja.