Opis problema

Imamo kavomat. Zanima nas, kolikšen je „v povprečju“ volumen ene skodelice kave (espresso), ki jo kavomat servira.

Statistična formulacija problema

- Slučajni eksperiment: (plačamo in) pritisnemo na gumb.
- Množica izidov Ω: skodelice s kavo (količina variira).
- Numerična kvalifikacija izida slučajnega eksperimenta?

  Slučajna spremenljivka $X \colon \Omega \to \mathbb{R}$,
  $X(\text{skodelica kave}) = \text{volumen}$.
- Kaj nas zanima?

  Matematično upanje $E(X)$.

Kako bomo ocenili pričakovano vrednost?

S pomočjo vzorca neke velikosti, ki jo označimo n.

1. Pritisnemo gumb, dobimo skodelico kave ω_1.
   Pišemo $X_1(\text{vzorec}) = X(\omega_1)$.
2. Pritisnemo gumb, dobimo skodelico kave ω_2.
   Pišemo $X_2(\text{vzorec}) = X(\omega_2)$.

   ...

n. Pritisnemo gumb, dobimo skodelico kave ω_n.
3. Pišemo $X_n(\text{vzorec}) = X(\omega_n)$.

Formalno gledano smo dobili slučajne spremenljivke $X_1, X_2, \ldots, X_n$, ki so vse definirane na prostoru vzorcev velikosti n.

Kako bomo ocenili pričakovano vrednost? Z vzorčnim povprečjem!

Vzorčno povprečje ...

... je spet slučajna spremenljivka. Pišemo $\overline{X}\colon \{\text{vzorci}\} \to \mathbb{R}$,

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

Cenilka

Slučajni spremenljivki, ki na podlagi vzorca ocenjuje neki parameter, ki nas zanima, pravimo cenilka.

Nepristranskost

Cenilka $T\colon \{\text{vzorci}\} \to \mathbb{R}$ je nepristranska, če je matematično upanje $E(T)$ enako parametru, ki ga ocenjujemo.

Ali je vzorčno povprečje $\overline{X}$ nepristranska cenilka za matematično upanje?

Oglejmo si porazdelitev slučajne spremenljivke X_1

Naj bo J interval (recimo $J = (a, b)$ ali $J = (a, b]$, kjer $a \in \mathbb{R}$ ali $a = -\infty$ ter $b \in \mathbb{R}$). Zanima nas $P(X_1 \in J)$.

Na katerem verjetnostnem prostoru je že definirana X_1?

Na prostoru vseh vzorcev velikosti n, torej

$$\mathbb{S} = \{(\omega_1, \omega_2, \ldots, \omega_n) \,|\, \omega_i \in \Omega\} = \Omega^n.$$

Ker je $X_1(\omega_1, \omega_2, \ldots, \omega_n) = X(\omega_1)$, je $P(X_1 \in J) = P(X \in J)$ oziroma, precizneje,

$$P(X_1 \in J) = P\Big(\{(\omega_1, \omega_2, \ldots, \omega_n) \,|\, X_1(\omega_1, \omega_2, \ldots, \omega_n) \in J\}\Big).$$

Gre torej za verjetnost dogodka $X^{-1}(J) \times \Omega \times \Omega \times \cdots \times \Omega$.

Z uporabo produktne verjetnosti sledi

$$P_{\mathbb{S}}(X^{-1}(J) \times \Omega \times \Omega \times \ldots \times \Omega) = P_\Omega(X^{-1}(J)) \cdot P_\Omega(\Omega) \ldots P_\Omega(\Omega).$$

V vsakem primeru torej velja:
$$P_{\mathbb{S}}(X_1 \in J) = P_{\Omega}(X \in J)$$
za vse intervale J.

Kaj to pomeni?

Slučajna spremenljivka X_1 je enako porazdeljena kot slučajna spremenljivka X. Posebej, velja $E(X_1) = E(X)$.

Z enakim premislekom vidimo, da so $X_1, X_2, \ldots, X_n$ vse enako porazdeljene in vse porazdeljene kot X.

Posledično:

$$E(\overline{X}) = E\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) = \frac{E(X_1)+E(X_2)+\cdots+E(X_n)}{n} = E(X).$$

Oglejmo si še cenilko

$$T = \frac{X_1+X_3+X_5+\cdots+X_{n \text{ ali } n-1}}{n/2 \text{ ali } (n+1)/2} = \frac{X_1 + X_3 + \cdots + X_{2\cdot\lceil n/2\rceil+1}}{\lceil n/2\rceil}.$$

Tudi T je nepristranska.

Katera je boljša, $\overline{X}$ ali T?

Definicija

Naj bosta T in U dve nepristranski cenilki za „parameter“ $\vartheta \in \mathbb{R}$, ki je povezan z opazovano slučajno spremenljivko X. Tedaj ima U enakomerno manjšo disperzijo od T, če velja

$$D(U) \leqslant D(T)$$

za vse možne porazdelitve za X.

Diskurz: disperzija neodvisnih slučajnih spremenljivk

Neodvisnost slučajnih spremenljivk

- X in Y sta neodvisni s.s., če za vse (primerne) A in B velja $P(X \in A \text{ in } Y \in B) = P(X \in A) \cdot P(Y \in B)$.
- Če sta X in Y neodvisni, velja $E(X \cdot Y) = E(X) \cdot E(Y)$.
- $X_1, X_2, \ldots, X_n$ so parome neodvisne, če sta X_i in X_j neodvisni za vsak par $i \neq j$.

Če so X_1, X_2, …, X_n paroma neodvisne, velja

$$D(X_1 + X_2 + \cdots + X_n) = D(X_1) + \cdots + D(X_n).$$

Dokaz je dovolj napraviti za dve:

$$D(X+Y) = E((X+Y)^2) - E(X+Y)^2$$
$$= E(X^2) + 2E(X \cdot Y) + E(Y^2) - E(X)^2 - 2E(X) \cdot E(Y) - E(Y)^2$$
$$= D(X) + D(Y).$$

$$T = \frac{X_1 + X_3 + \cdots + X_{2\cdot\lceil n/2\rceil+1}}{\lceil n/2\rceil}.$$

Podobno kot zgoraj računamo:

$P(X_1 \in I \text{ in } X_2 \in J) = P_{\mathbb{S}}(X^{-1}(I) \times X^{-1}(J) \times \Omega \times \cdots \times \Omega)$
$= P_\Omega(X \in I) \cdot P_\Omega(X \in J) = (\text{že vemo}) P_{\mathbb{S}}(X_1 \in I) \cdot P_{\mathbb{S}}(X_2 \in J)$.
Sklenemo: $P_{\mathbb{S}}(X_1 \in I \text{ in } X_2 \in J) = P_{\mathbb{S}}(X_1 \in I) \cdot P_{\mathbb{S}}(X_2 \in J)$.

Torej sta X_1 in X_2 neodvisni. Podobno za katerikoli par $i \neq j$ dokažemo, da sta X_i in X_j neodvisni.

Končno sledi:

- $D(\overline{X}) = D(\frac{X_1+\cdots+X_n}{n}) = \frac{1}{n^2} D(X_1 + \ldots + X_n) = \frac{1}{n^2} \cdot nD(X) = \frac{D(X)}{n}$.
- Podobno: $D(T) = D(X)/\lceil n/2\rceil$.

Opis problema

Zanima nas delež tistih uporabnikov mobilne telefonije ponudnika A, ki so zadovoljni s storitvami, ki jih koristijo (in plačujejo).

Statistična formulacija problema

- Slučajni eksperiment: izberemo (ali naletimo na) naključnega uporabnika.
- Množica izidov Ω: vsi uporabniki.
- Numerična kvalifikacija izida slučajnega eksperimenta?

  Slučajna spremenljivka $X : \Omega \to \mathbb{R}$,

  $$X(\text{uporabnik}) = \begin{cases} 1, & \text{če je zadovoljen,} \\ 0, & \text{če je nezadovoljen.} \end{cases}$$
- Kaj nas zanima?

  Matematično upanje $E(X)$!

Bernoullijeva slučajna spremenljivka ...

... je slučajna spremenljivka X z vrednostma 1 in 0.

- Pišemo $p = P(X = 1)$. Tedaj avtomatično $1 - p = P(X = 0)$.
- $E(X) = 1 \cdot p + 0 \cdot (1 - p) = p$.
- Bernoullijeva spremenljivka je porazdeljena kot $\mathrm{Bin}(1, p)$.

Ponovimo

Če izide nekega slučajnega eksperimenta lahko razdelimo na „ugodne“ in „neugodne“, je delež ugodnih izidov enak matematičnemu upanju Bernoullijeve slučajne spremenlivke X, za katero

$$X(\text{ugoden izid}) = 1 \text{ in } X(\text{neugoden izid}) = 0.$$

Vrnimo se k našim (ne)zadovoljnim uporabnikom mobilnih storitev.

Kako bomo ocenili matematično upanje?

S pomočjo vzorca neke velikosti, ki jo označimo n.
Kaj najprej naredimo?
Naključno izberemo n različnih uporabnikov $\omega_1, \omega_2, \ldots, \omega_n$.

1. Beležimo $X_1(\text{vzorec}) = X(\omega_1)$ (ena ali nič).
2. Beležimo $X_2(\text{vzorec}) = X(\omega_2)$ (ena ali nič).

   ...

n. Beležimo $X_n(\text{vzorec}) = X(\omega_n)$ (ena ali nič).

Formalno gledano smo dobili slučajne spremenljivke $X_1, X_2, \ldots, X_n$, ki so vse definirane na prostoru vzorcev velikosti n.

Kako bomo ocenili matematično upanje? Z vzorčnim povprečjem!

Vzorčenje: 3. primer

Opis problema

Eden od naših zaposlenih strankam nudi podporo na domu. Zanima nas delež njegovih storitev, ki so ocenjene zadovoljivo. (Poudarek je na storitvah: k isti stranki gre lahko pogosto.)

Kako bomo delež ocenili?

Izbrali bomo vzorec n njegovih storitev $\omega_1, \omega_2, \ldots, \omega_n$. Slučajni eksperiment je tu $X\colon \Omega \to \{0, 1\}$, kjer je Ω množica vseh možnih storitev, ki jih mora opraviti. Gre za Bernoullijevo slučajno spremenljivko. Oceniti želimo $E(X)$. Podobno kot prej:

1. $X_1(\text{vzorec}) = X(\omega_1)$ (ena ali nič),
2. $X_2(\text{vzorec}) = X(\omega_2)$ (ena ali nič),

   ...

n. $X_n(\text{vzorec}) = X(\omega_n)$ (ena ali nič).

Cenilka za delež je vzorčno povprečje $\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$.

Podobnosti / razlike med primeri?

Poizkusimo izračunati kar verjetnostne porazdelitve.

Zadovoljstvo uporabnikov mobilnih storitev

- Recimo, da je vseh uporabnikov natanko N.
- Predpostavimo, da je od teh $M = N \cdot p$ zadovoljnih.
- Kolikšna je verjetnost, da je v vzorcu n različnih uporabnikov natanko k zadovoljnih (tu je $k \leqslant n$ in $k \leqslant M$)?

  Gre za $\dfrac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}}$.
- To je hipergeometrična porazdelitev.
- Možne vrednosti vzorčnega povprečja so $0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{\min\{n,M\}}{n}$ in $P(\overline{X} = \frac{k}{n}) = \frac{\binom{M}{k}\cdot\binom{N-M}{n-k}}{\binom{N}{n}}$.

Podobnosti / razlike med primeri, drugič

Zadovoljstvo s storitvami našega zaposlenega

- Recimo, da je delež storitev, ki so ocenjene kot zadovoljive, natanko p.
- Kolikšna je verjetnost, da bo v vzorcu n storitev natanko k ocenjenih kot zadovoljivih (tu je $k \leqslant n$)?

  Gre za $\displaystyle\binom{n}{k} p^k (1-p)^{n-k}$.
- To je binomska porazdelitev.
- Možne vrednosti vzorčnega povprečja so $0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n} = 1$ in $P(\overline{X} = \frac{k}{n}) = \binom{n}{k} p^k (1-p)^{n-k}$.

Kaj pa kavomat?

- Ni tako preprosto ☹
- Vsekakor gre za zvezno porazdeljeno slučajno spremenljivko.
- Privzamemo, da $P(\overline{X} \in (a,b)) = P(X_1 + \cdots + X_n \in (na, nb)) = \int_a^b g(t)\mathrm{d}t$, kjer je $g(t)$ porazdelitvena gostota vsote $X_1 + X_2 + \cdots + X_n$.
- Velja ☹
$$g(t) = \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} f(x_1) \cdots f(x_{n-1}) f(t - \sum_{i=1}^{n-1} x_i) \mathrm{d}x_1 x_2 \ldots x_{n-1}.$$
- Če je X normalno porazdeljena slučajna spremenljivka, $X \sim N(\mu, \sigma)$, je $X_1 + X_2 + \cdots + X_n \sim N(n\mu, \sigma\sqrt{n})$ in $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$. ☺

Vzorčenje s ponavljanjem, brez ponavljanja

Komponente slučajnega vzorca

Slučajne spremenljivke $X_1, X_2, \ldots X_n$, za katere velja

$$X_i(\text{vzorec}) = X(\omega_i),$$

imenujemo komponente slučajnega vzorca.

Vzorčenje s ponavljanjem

Kadar so komponente slučajnega vzorca neodvisne druga od druge (izid ω_i je dobljen neodvisno od izidov $\omega_1, \omega_2, \ldots, \omega_{i-1}$ in je lahko enak kateremu od njih), pravimo, da gre za vzorčenje s ponavljanjem).

Vzorčenje brez ponavljanja

Kadar je Ω končna množica in zahtevamo, da so izidi $\omega_1, \ldots, \omega_n$ vsi različni, pravimo, da gre za vzorčenje brez ponavljanja).

Disperzija neodvisnih slučajnih spremenljivk

Neodvisnost slučajnih spremenljivk

- X in Y sta neodvisni s.s., če za vse A in B velja
  $P(X \in A \text{ in } Y \in B) = P(X \in A) \cdot P(Y \in B)$.
- $X_1, \ldots, X_n$ so neodvisne, če velja
  $P(X_1 \in A_1 \text{ in } \ldots X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n)$.

Disperzija neodvisnih spremenljivk

Če so X_1, X_2, ..., X_n neodvisne, velja
$D(X_1 + X_2 + \cdots + X_n) = D(X_1) + \cdots + D(X_n)$.

Disperzija cenilke za matematično upanje

$D(\overline{X})$ za vzorčenje s ponavljanjem

$$D(\overline{X}) = \tfrac{1}{n^2}\left(D(X_1) + \cdots + D(X_n)\right) = \tfrac{n}{n^2} D(X) = \frac{D(X)}{n}.$$

Če pišemo $\sigma_X = \sqrt{D(X)}$, je $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$.

$D(\overline{X})$ za vzorčenje brez ponavljanja

$$D(\overline{X}) = \frac{D(X)}{n}\left(1 - \frac{n-1}{N-1}\right) = \frac{\sigma_X^2}{n}\left(1 - \frac{n-1}{N-1}\right),$$

kjer je N število elementov množice Ω.

Opis problema

Eden od naših zaposlenih strankam nudi podporo na domu. Zanima nas delež njegovih storitev, ki so ocenjene zadovoljivo. (Poudarek je na storitvah: k isti stranki gre lahko pogosto.)

Statistična formulacija problema

- Slučajni eksperiment: „izberemo“ storitev (zaposleni jo stori).
- Množica izidov Ω: vse potencialno mogoče storitve.
- Numerična kvalifikacija izida slučajnega eksperimenta?

  Slučajna spremenljivka $X : \Omega \to \mathbb{R}$,

  $$X(\text{storitev}) = \,?\begin{cases} 1, & \text{če je storitev zadovoljiva,} \\ 0, & \text{če je nezadovoljiva.} \end{cases}$$

- Kaj nas zanima?

  Matematično upanje $E(X)$!

Bernoullijeva slučajna spremenljivka ...

... je slučajna spremenljivka X z vrednostma 1 in 0.

- Pišemo $p = P(X = 1)$. Tedaj $1 - p = P(X = 0)$ in $E(X) = p$.
- Označimo $X \sim B(1, p)$.

Povzemimo

Če izide nekega slučajnega eksperimenta lahko razdelimo na „ugodne“ in „neugodne“, je delež ugodnih izidov enak mat. upanju Bernoullijeve slučajne spremenljivke X, za katero:

$$X(\text{ugoden izid}) = 1 \text{ in } X(\text{neugoden izid}) = 0.$$

Modeliranje je določanje možnih porazdelitev za dani slučajni eksperiment. Če ima naš eksperiment Bernoullijevo porazdelitev (za neki neznani parameter $p \in (0, 1)$), pravimo, da gre za Bernoullijev model.

Vrnimo se k (ne)zadovoljivim storitvam zaposlenega v podpori.

Kako bomo ocenili matematično upanje?

S pomočjo vzorca neke velikosti, ki jo označimo n.
Kako? Naključno izberemo n storitev $\omega_1, \omega_2, \ldots, \omega_n$.

1. Beležimo $X_1(\text{vzorec}) = X(\omega_1)$ (ena ali nič).
2. Beležimo $X_2(\text{vzorec}) = X(\omega_2)$ (ena ali nič).

   . . .

n. Beležimo $X_n(\text{vzorec}) = X(\omega_n)$ (ena ali nič).

Formalno gledano smo dobili slučajne spremenljivke $X_1, X_2, \ldots, X_n$, ki so vse definirane na prostoru vzorcev velikosti n.

Kako bomo ocenili matematično upanje $E(X) = p$?
Z vzorčnim povprečjem!

Problem „točkovne ocene“

- Recimo, da napravimo vzorec 10 storitev.
- Naj bo k število zadovoljivih storitev.
- Ocena za iskani delež je torej $\frac{k}{10}$.

V razmislek

Denimo, da dejanski p znaša 0.75.
Ali imamo kakšne možnosti, da bomo z oceno oblike $\frac{k}{10}$ točni??

Rešitev?

„Intervalska“ ocena: $p \in [k/10 - \text{nekaj}, k/10 + \text{nekaj}]$.
Predlogi? Poizkusimo z intervalsko oceno

$$„p \in [k/10 - 1/10, k/10 + 1/10].“$$

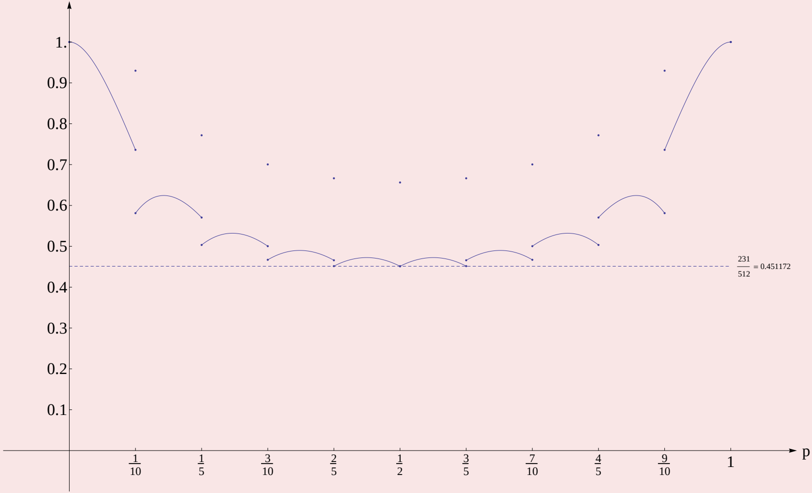
Kaj nas zanima?

Kako „točna“ je naša intervalska ocena!

Stopnja zaupanja?

k	*interval*	$p = 0.1$	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 0.9$
0	$[-.1, .1]$	$\boxed{0.349}$	0.056	0.000	0.000	0.000
1	$[0, .2]$	$\boxed{0.387}$	0.188	0.010	0.000	0.000
2	$[.1, .3]$	$\boxed{0.194}$	$\boxed{0.282}$	0.044	0.000	0.000
3	$[.2, .4]$	0.057	$\boxed{0.250}$	0.117	0.003	0.000
4	$[.3, .5]$	0.011	0.146	$\boxed{0.205}$	0.016	0.000
5	$[.4, .6]$	0.001	0.058	$\boxed{0.246}$	0.058	0.001
6	$[.5, .7]$	0.000	0.016	$\boxed{0.205}$	0.146	0.011
7	$[.6, .8]$	0.000	0.003	0.117	$\boxed{0.250}$	0.057
8	$[.7, .9]$	0.000	0.000	0.044	$\boxed{0.282}$	$\boxed{0.194}$
9	$[.8, 1]$	0.000	0.000	0.010	0.188	$\boxed{0.387}$
10	$[.9, 1.1]$	0.000	0.000	0.000	0.056	$\boxed{0.349}$
	pokritost	0.930	0.532	0.656	0.532	0.930

Graf pokritosti



Interval zaupanja: uvodna definicija

Uvodna definicija

Zanima nas mat. upanje $E(X)$ naše slučajne spremenljivke.
Naj bo velikost vzorca n in naj bo $\beta \in (0,1)$ primerno (dovolj „veliko“) število.
Definicija: Interval zaupanja za $E(X)$ stopnje zaupanja β za vzorce velikosti n sestavljata taki merljivi funkciji vzorca

$$L\colon \mathbb{R}^n \to \mathbb{R}, \qquad U\colon \mathbb{R}^n \to \mathbb{R},$$

za kateri velja

$$P\Big(L(X_1, X_2, \ldots, X_n) \leqslant E(X) \leqslant U(X_1, X_2, \ldots, X_n)\Big) \geqslant \beta$$

za vse „možne porazdelitve“ eksperimenta X.
Pravimo tudi, da je interval zaupanja prireditev

$$\vec{X} \mapsto [L(\vec{X}), U(\vec{X})].$$

V tem smislu je interval zaupanja slučajni interval.

Kakšen je naš naivni interval zaupanja $[k/10 - 1/10, k/10 + 1/10]$?

Ima stopnjo zaupanja približno 0.45.

Interval zaupanja za μ v normalnem modelu z znanim σ

Izrek

Naj velja $P(N(0,1) \in (a,b)) = P(Z \in (a,b)) = \beta$.
Slučajni interval (za vzorec velikosti n)

$$\vec{X} = (X_1, \ldots, X_n) \mapsto \left[\overline{X} - b \cdot \sigma/\sqrt{n}, \overline{X} - a \cdot \sigma/\sqrt{n}\right]$$

je interval zaupanja za μ stopnje zaupanja (natanko) β.

Zgornji interval je najkrajši, če je
$-a = b = z_{(1-\beta)/2} = \Phi^{-1}((1+\beta)/2)$, zgornji $(1-\beta)/2$-percentil standardne normalne porazdelitve.

Torej $P(\mu \in [\overline{X} - z_{(1-\beta)/2} \cdot \sigma/\sqrt{n}, \overline{X} + z_{(1-\beta)/2} \cdot \sigma/\sqrt{n}]) = \beta$.