Statistika

6. predavanje

Barbara Boldin

Fakulteta za matematiko, naravoslovje in informacijske tehnologije
Univerza na Primorskem

Kovarianca in korelacija

Varianca $Var(X)$ je merilo razpršenosti slučajne spremenljivke X okoli pričakovane vrednosti $\mu = E(X)$.

Kovarianca in **korelacija** merita povezanost dveh slučajnih spremenljivk.

Npr:: Ali je glasnost oglašanja čričkov povezana s temperaturo? Ali je zaslužek povezan z izobrazbo?

Naj bosta X in Y slučajni spremenljivki, $\mu_X = E(X)$ in $\mu_Y = E(Y)$.
Kovarianca je

$$Cov(X, Y) = E\Big((X - \mu_X)(Y - \mu_Y)\Big) = E(XY) - E(X)E(Y)$$

Kadar sta X in Y neodvisni slučajni spremeljivki, je $Cov(X, Y) = 0$.

Korelacija med X in Y je

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

(privzamemo $Var(X) \neq 0$ in $Var(Y) \neq 0$).

Zakaj bi uporabljali bolj zapleten izraz $\rho(X, Y)$ namesto $Cov(X, Y)$? Vrednost $\rho(X, Y)$ je neodvisna od izbire enot slučajnih spremenljivk X in Y!

Kovarianca in korelacija

Varianca $Var(X)$ je merilo razpršenosti slučajne spremenljivke X okoli pričakovane vrednosti $\mu = E(X)$.

Kovarianca in **korelacija** merita povezanost dveh slučajnih spremenljivk.

Npr:: Ali je glasnost oglašanja čričkov povezana s temperaturo? Ali je zaslužek povezan z izobrazbo?

Naj bosta X in Y slučajni spremenljivki, $\mu_X = E(X)$ in $\mu_Y = E(Y)$.
Kovarianca je

$$Cov(X, Y) = E\Big((X - \mu_X)(Y - \mu_Y)\Big) = E(XY) - E(X)E(Y)$$

Kadar sta X in Y neodvisni slučajni spremeljivki, je $Cov(X, Y) = 0$.

Korelacija med X in Y je

$$\rho(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$$

(privzamemo $Var(X) \neq 0$ in $Var(Y) \neq 0$).

Zakaj bi uporabljali bolj zapleten izraz $\rho(X, Y)$ namesto $Cov(X, Y)$? Vrednost $\rho(X, Y)$ je neodvisna od izbire enot slučajnih spremenljivk X in Y!

Primer. Naj bo slučajna spremenljivka X podana s funkcijo verjetnosti

k	$P(X = k)$
-1	0.2
0	0.4
1	0.4

in $Y = |X| + 1$. Izračunajmo $Cov(X, Y)$.

Funkcija verjetnosti za spremenljivko Y je

k	$P(Y = k)$
1	0.4
2	0.6

saj je

$$P(Y = 1) = P(X = 0) = 0.4,$$
$$P(Y = 2) = P(X = 1) + P(X = -1) = 0.6.$$

Torej

$$E(Y) = 1 \cdot 0.4 + 2 \cdot 0.6 = 1.6$$

Primer. Naj bo slučajna spremenljivka X podana s funkcijo verjetnosti

k	$P(X = k)$
-1	0.2
0	0.4
1	0.4

in $Y = |X| + 1$. Izračunajmo $Cov(X, Y)$.

Funkcija verjetnosti za spremenljivko Y je

k	$P(Y = k)$
1	0.4
2	0.6

saj je

$$P(Y = 1) = P(X = 0) = 0.4,$$
$$P(Y = 2) = P(X = 1) + P(X = -1) = 0.6.$$

Torej

$$E(Y) = 1 \cdot 0.4 + 2 \cdot 0.6 = 1.6$$

Za spremenljivko XY je funkcija verjetnosti

k	$P(XY = k)$
-2	0.2
0	0.4
2	0.4

saj je

$$P(XY = -2) = P(X = -1) = 0.2,$$
$$P(XY = 0) = P(X = 0) = 0.4$$
$$P(XY = 2) = P(X = 1) = 0.4.$$

Torej imamo $E(XY) = 0.4$. Ker je $E(X) = 0.2$ sledi

$$Cov(X, Y) = E(XY) - E(X)E(Y) = 0.08$$

Za spremenljivko XY je funkcija verjetnosti

k	$P(XY = k)$
-2	0.2
0	0.4
2	0.4

saj je

$$P(XY = -2) = P(X = -1) = 0.2,$$
$$P(XY = 0) = P(X = 0) = 0.4$$
$$P(XY = 2) = P(X = 1) = 0.4.$$

Torej imamo $E(XY) = 0.4$. Ker je $E(X) = 0.2$ sledi

$$Cov(X, Y) = E(XY) - E(X)E(Y) = 0.08$$

Vzorčenje

Vzorčenje je način pridobivanja informacij o (načeloma veliki) populaciji s preučevanjem dela populacije (vzorca).

Npr.:

- javnomnenjske ankete pred volitvami,
- v kmetijstvu kvaliteto pridelka ocenjujemo na podlagi kvalitete vzorca,
- učinkovitost novega cepiva ocenjujemo na podlagi učinkovitosti na testni skupini,
- itd.

Vzorčenje lahko poteka:

- po nekem pravilu (npr.: prvih 1000 v telefonskem imeniku),
- naključno

Naključno vzorčenje ima več prednosti: omogoča oceno napake, kar pomeni, da velikost vzorcev pri raziskavi lahko izberemo tako, da bo napaka pod določeno vrednostjo (načrtovanje raziskave!)

Točkasto ocenjevanje parametrov

Denimo, da je celotna populacija velikosti N, vrednosti preučevane slučajne spremenljivke X v populaciji pa $x_1, x_2, \ldots, x_N$.

Želimo oceniti naslednje statistike:

- populacijsko povprečje (aritmetična sredina)

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

- vsota

$$\tau = x_1 + x_2 + \ldots + x_N$$

Vsota enot ima smisel predvsem za binarne spremenljivke, ki beležijo pristotnost/odsotnost neke lastnosti. Če 1/0 pomeni prisotnost/odsotnost lastnosti, potem je τ število enot z dano lastnostjo.

- populacijska varianca

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \mu^2$$

- populacijski standardni odklon σ

Iz populacije naključno vzamemo vzorec velikosti n. Ali je aritmetična sredina (vsota, varianca, standardni odklon) vzorca enaka populacijski aritmetični sredini (vsoti, varianci, standardnem odklonu)? Ne!

Primer. V populaciji velikosti $N = 10$ so vrednosti neke slučajne spremenljivke

$$1, 2, 0, 3, 6, 5, 9, 7, 4, 2,$$

torej $\mu = 3.9$. Izberemo tri naključne vzorce velikosti $n = 3$ in dobimo:

- vzorec 1: 1, 5, 0 ima sredino 2,
- vzorec 2: 3, 0, 9 ima sredino 4,
- vzorec 3: 4, 6, 5 ima sredino 5.

Kako dober približek za μ (oz. τ, σ^2, σ) dobimo z vzorčenjem?

Denimo, da iz populacije naključno izberemo vzorec velikosti n. Ker je vzorec naključen, so vzorčne vrednosti

$$X_1, X_2, \ldots, X_n$$

slučajne spremenljivke, torej je slučajno tudi povprečje

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Porazdelitev $\bar{X}$ imenujemo **vzorčna porazdelitev**.

Izkaže se, da je

$$E(\bar{X}) = \mu$$

V dolgoročnem povprečju je torej ocena aritmetične sredine populacije enaka μ.

Rečemo, da je **vzorčenje nepristransko**, vzorčno povprečje pa nepristranska ocena populacijskega povprečja.

Denimo, da iz populacije naključno izberemo vzorec velikosti n. Ker je vzorec naključen, so vzorčne vrednosti

$$X_1, X_2, \ldots, X_n$$

slučajne spremenljivke, torej je slučajno tudi povprečje

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Porazdelitev $\bar{X}$ imenujemo **vzorčna porazdelitev**.

Izkaže se, da je

$$\boxed{E(\bar{X}) = \mu}$$

V dolgoročnem povprečju je torej ocena aritmetične sredine populacije enaka μ.

Rečemo, da je **vzorčenje nepristransko**, vzorčno povprečje pa nepristranska ocena populacijskega povprečja.

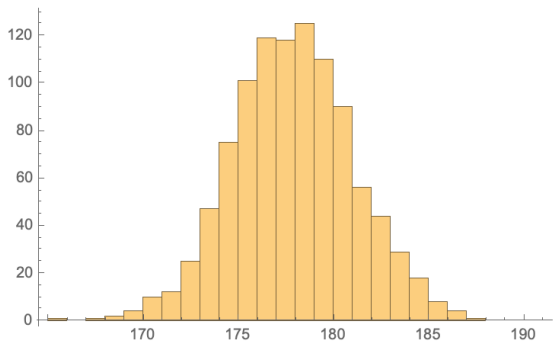
Kakšna pa je varianca $\bar{X}$? Izkaže se, da je

$$Var(\bar{X}) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

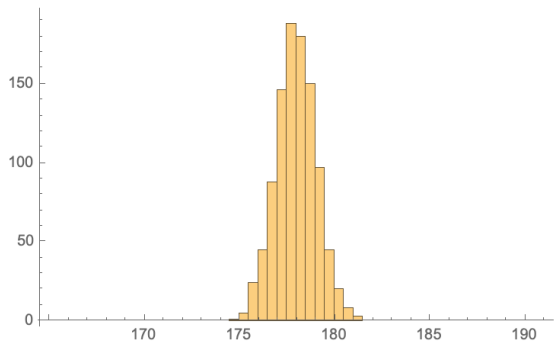
Faktor $\frac{n-1}{N-1}$ je **popravek končne populacije**. Če je $n \ll N$, lahko popravek zanemarimo in

$$Var(\bar{X}) \approx \frac{\sigma^2}{n}$$

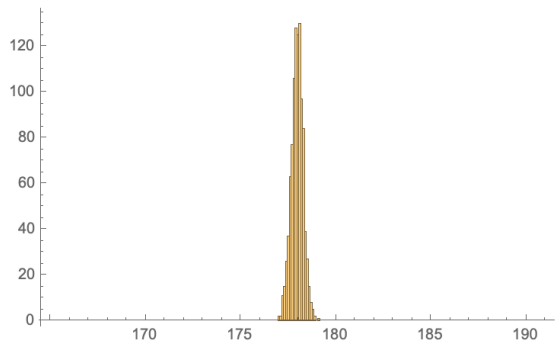
Odvisnost variance vzorčne porazdelitve od velikosti vzorca prikazuje naslednji poskus: višina odraslih moških v populaciji je porazdeljena normalno z $\mu = 178$ cm in $\sigma = 10$ cm. Iz populacije vzamemo vzorec velikosti n in izračunamo povprečje vzorca. Poskus ponovimo 1000-krat in s histogramom prikažemo porazdelitev vzorčnih aritmetičnih sredin.



Histogram povprečij 1000 vzorcev, $n = 10$



Histogram povprečij 1000 vzorcev, $n = 100$



Histogram povprečij 1000 vzorcev, $n = 1000$

Oceno za $Var(\bar{X})$ smo dobili s pomočjo populacijske variance σ^2, ki pa v splošnem ni poznana. Kako bi dobili oceno za varianco populacije?

Prvi predlog bi bil

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Izkaže se, da ta ocena ni nepristranska ($E(\hat{\sigma}^2) \neq \sigma^2$), velja

$$E(\hat{\sigma}^2) = \sigma^2 \frac{N}{N-1}\frac{n-1}{n}.$$

Ker je $\frac{N}{N-1}\frac{n-1}{n} < 1$, varianco vedno **podcenimo**. Nepristranska ocena za varianco je $\frac{N-1}{N}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Za velike populacije N je **nepristranska ocena variance**

$$\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Nepristranska ocena za varianco vzorčne porazdelitve, $Var(\bar{X})$, je

$$\boxed{s^2_{\bar{X}} = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$$

Oceno za $Var(\bar{X})$ smo dobili s pomočjo populacijske variance σ^2, ki pa v splošnem ni poznana. Kako bi dobili oceno za varianco populacije? Prvi predlog bi bil

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Izkaže se, da ta ocena ni nepristranska ($E(\hat{\sigma}^2) \neq \sigma^2$), velja

$$E(\hat{\sigma}^2) = \sigma^2 \frac{N}{N-1}\frac{n-1}{n}.$$

Ker je $\frac{N}{N-1}\frac{n-1}{n} < 1$, varianco vedno **podcenimo**. Nepristranska ocena za varianco je $\frac{N-1}{N}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Za velike populacije N je **nepristranska ocena variance**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Nepristranska ocena za varianco vzorčne porazdelitve, $Var(\bar{X})$, je

$$s^2_{\bar{X}} = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$$

Oceno za $Var(\bar{X})$ smo dobili s pomočjo populacijske variance σ^2, ki pa v splošnem ni poznana. Kako bi dobili oceno za varianco populacije? Prvi predlog bi bil

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Izkaže se, da ta ocena ni nepristranska ($E(\hat{\sigma}^2) \neq \sigma^2$), velja

$$E(\hat{\sigma}^2) = \sigma^2 \frac{N}{N-1}\frac{n-1}{n}.$$

Ker je $\frac{N}{N-1}\frac{n-1}{n} < 1$, varianco vedno **podcenimo**. Nepristranska ocena za varianco je $\frac{N-1}{N}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Za velike populacije N je **nepristranska ocena variance**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Nepristranska ocena za varianco vzorčne porazdelitve, $Var(\bar{X})$, je

$$s^2_{\bar{X}} = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$$

Nepristranske ocene populacijskih parametrov μ, τ in σ^2 zberimo v tabeli:

Populacijski parameter	Nepristranska ocena
μ	$\frac{1}{n}\sum_{i=1}^{n} X_i$
τ	$\frac{N}{n}\sum_{i=1}^{n} X_i$
σ^2	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$

Primer. Naključno smo izbrali devet štiriletnih deklic in izmerili njihovo višino. Dobili smo naslednje podatke (v cm)

$$101, 91, 93, 103, 91, 101, 103, 95, 95$$

Nepristranska ocena za višino štiriletnic v celotni populaciji je

$$\bar{x} = \frac{1}{9}\big(101 + 91 + \ldots + 95\big) = 97cm$$

Kolikšna je nepristranska ocena za standardni odklon v populaciji?

$$s^2 = \frac{1}{8}\big((101 - 97)^2 + (91 - 97)^2 + \ldots + (95 - 97)^2\big) = 25cm^2$$

torej

$$s = 5cm$$

Primer. Naključno smo izbrali devet štiriletnih deklic in izmerili njihovo višino. Dobili smo naslednje podatke (v cm)

$$101, 91, 93, 103, 91, 101, 103, 95, 95$$

Nepristranska ocena za višino štiriletnic v celotni populaciji je

$$\bar{x} = \frac{1}{9}\big(101 + 91 + \ldots + 95\big) = 97cm$$

Kolikšna je nepristranska ocena za standardni odklon v populaciji?

$$s^2 = \frac{1}{8}\big((101 - 97)^2 + (91 - 97)^2 + \ldots + (95 - 97)^2\big) = 25cm^2$$

torej

$$s = 5cm$$

Primer. Naključno smo izbrali devet štiriletnih deklic in izmerili njihovo višino. Dobili smo naslednje podatke (v cm)

$$101, 91, 93, 103, 91, 101, 103, 95, 95$$

Nepristranska ocena za višino štiriletnic v celotni populaciji je

$$\bar{x} = \frac{1}{9}\Big(101 + 91 + \ldots + 95\Big) = 97cm$$

Kolikšna je nepristranska ocena za standardni odklon v populaciji?

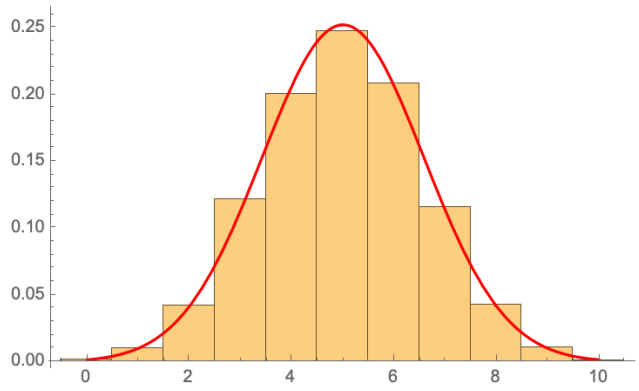
$$s^2 = \frac{1}{8}\Big((101 - 97)^2 + (91 - 97)^2 + \ldots + (95 - 97)^2\Big) = 25cm^2$$

torej

$$s = 5cm$$

Aproksimacija vzorčne porazdelitve z normalno

Ob velikem številu vzorcev histogrami vzorčne porazdelitve $\bar{X}$ zgledajo približno normalno, četudi spremenljivka X ni porazdeljena normalno.



$X \sim B(10, 0.5)$

Da bi lahko odgovorili na vprašanje

kako tipičen je nek slučajni vzorec?

porazdelitev $\bar{X}$ aproksimiramo z normalno slučajno spremenljivko: če je μ pričakovana vrednost spremenljivke X in σ^2 njena varianca, potem $\bar{X}$ aproksimiramo z $N(\mu, \frac{\sigma^2}{n})$

Označimo z $\bar{X}_n$ vzorčno porazdelitev, če so vzorci velikosti n.
Če F označuje porazdelitveno funkcijo standarne normalne slučajne spremenljivke $N(0,1)$, potem t.i. centralni limitni izrek pove, da

$$P\left(\frac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z\right) \to F(z)$$

Z besedami: ko se velikost vzorca povečuje, je porazdelitvena funkcija spremenljivke

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

vse bližje porazdelitveni funkciji standardne normalne porazdelitve. Ker porazdelitveno funkcijo $N(0,1)$ poznamo (vsaj iz tabele), lahko sedaj ocenimo, kako tipičen je nek vzorec.

Primer. Denimo, da je višina v populaciji štiriletnih deklic normalno porazdeljena z $\mu = 100$ cm in $\sigma = 4$ cm. Iz populacije smo izbrali devet štiriletnih deklic, jim izmerili višino in dobili $\bar{x} = 97$ cm.

- Kako tipičen je ta vzorec, t.j. kakšen delež vseh vzorcev velikosti $n = 9$ ima povprečje vsaj 97 cm?

$$P(\bar{X} \geq 97) = P\left(\frac{\bar{X} - 100}{\frac{4}{3}} \geq \frac{97 - 100}{\frac{4}{3}}\right) = P(Z \geq -\frac{9}{4})$$

$$= 1 - P(Z < -\frac{9}{4}) = 1 - 0.0122 = 0.9878$$

Torej približno $98,8\%$ vzorcev velikost 9 ima povprečje vsaj 97 cm.

Primer. Denimo, da je višina v populaciji štiriletnih deklic normalno porazdeljena z $\mu = 100$ cm in $\sigma = 4$ cm. Iz populacije smo izbrali devet štiriletnih deklic, jim izmerili višino in dobili $\bar{x} = 97$ cm.

- Kako tipičen je ta vzorec, t.j. kakšen delež vseh vzorcev velikosti $n = 9$ ima povprečje vsaj 97 cm?

$$P(\bar{X} \geq 97) = P\left(\frac{\bar{X} - 100}{\frac{4}{3}} \geq \frac{97 - 100}{\frac{4}{3}}\right) = P(Z \geq -\frac{9}{4})$$
$$= 1 - P(Z < -\frac{9}{4}) = 1 - 0.0122 = 0.9878$$

Torej približno $98,8\%$ vzorcev velikost 9 ima povprečje vsaj 97 cm.

- Kakšen delež aritmetičnih sredin vseh vzorcev velikosti $n = 9$ se od populacijskega povprečja razlikuje vsaj za 3cm?

$$
\begin{aligned}
P\left(|\bar{X} - 100| \geq 3\right) &= P\left(\bar{X} \leq 97\right) + P\left(\bar{X} \geq 103\right) \\
&= 2 \cdot P\left(\bar{X} \leq 97\right) \ (\textit{zaradi simetrije}) \\
&= 2 \cdot P\left(\frac{\bar{X} - 100}{\frac{4}{3}} \leq \frac{97 - 100}{\frac{4}{3}}\right) \\
&= 2 \cdot P(Z \leq -\frac{9}{4}) \\
&= 0.0244
\end{aligned}
$$

Torej pri približno 2.4% vzorcev velikosti 9 se aritmetična sredina vzorca za vsaj 3 cm razlikuje od μ.

- Kakšen delež aritmetičnih sredin vseh vzorcev velikosti $n = 9$ se od populacijskega povprečja razlikuje vsaj za 3cm?

$$
\begin{aligned}
P\left(|\bar{X} - 100| \geq 3\right) &= P\left(\bar{X} \leq 97\right) + P\left(\bar{X} \geq 103\right) \\
&= 2 \cdot P\left(\bar{X} \leq 97\right) \ (\textit{zaradi simetrije}) \\
&= 2 \cdot P\left(\frac{\bar{X} - 100}{\frac{4}{3}} \leq \frac{97 - 100}{\frac{4}{3}}\right) \\
&= 2 \cdot P(Z \leq -\frac{9}{4}) \\
&= 0.0244
\end{aligned}
$$

Torej pri približno 2.4% vzorcev velikosti 9 se aritmetična sredina vzorca za vsaj 3 cm razlikuje od μ.

- Kako veliki naj bodo vzorci, da bo delež vzorcev, katerih sredina se bo od populacijskega povprečja razlikovala za največ 2 cm večji od 90 %?

Iščemo n, da bo

$$P(|\bar{X} - 100| \leq 2) = 0.9$$

Potem bo za vsak $m \geq n$ veljajo $P(|\bar{X} - 100| \leq 2) \geq 0.9$.

$$P(|\bar{X} - 100| \leq 2) = P(98 \leq \bar{X} \leq 102) = 1 - 2P(\bar{X} \leq 98) = 0.9$$

Torej

$$P(\bar{X} \leq 98) = 0.05.$$

Če označimo $Z = \frac{\bar{X}-100}{\frac{4}{\sqrt{n}}}$ in $z = \frac{98-100}{\frac{4}{\sqrt{n}}}$ imamo torej $P(Z \leq z) = 0.05$.
Iz tabele za $N(0,1)$ razberemo $z = -1.65$, torej

$$\frac{-2}{\frac{4}{\sqrt{n}}} = -1.65$$

in od tod $\sqrt{n} = 3.3$. Za $n \geq 11$ bo torej delež vzorcev, katerih sredina se bo od populacijskega povprečja razlikovala za manj kot 2 cm večji od 90 %.

- Kako veliki naj bodo vzorci, da bo delež vzorcev, katerih sredina se bo od populacijskega povprečja razlikovala za največ 2 cm večji od 90 %?

  Iščemo n, da bo
  $$P(|\bar{X} - 100| \leq 2) = 0.9$$
  Potem bo za vsak $m \geq n$ veljajo $P(|\bar{X} - 100| \leq 2) \geq 0.9$.

  $$P(|\bar{X} - 100| \leq 2) = P(98 \leq \bar{X} \leq 102) = 1 - 2P(\bar{X} \leq 98) = 0.9$$

  Torej
  $$P(\bar{X} \leq 98) = 0.05.$$
  Če označimo $Z = \frac{\bar{X}-100}{\frac{4}{\sqrt{n}}}$ in $z = \frac{98-100}{\frac{4}{\sqrt{n}}}$ imamo torej $P(Z \leq z) = 0.05$.
  Iz tabele za $N(0,1)$ razberemo $z = -1.65$, torej
  $$\frac{-2}{\frac{4}{\sqrt{n}}} = -1.65$$
  in od tod $\sqrt{n} = 3.3$. Za $n \geq 11$ bo torej delež vzorcev, katerih sredina se bo od populacijskega povprečja razlikovala za manj kot 2 cm večji od 90 %.

Primer. Iz populacije velikosti 1000 naključno izberemo 15 ljudi in jih prosimo, da rešijo IQ test. Dobimo nasledje rezultate

117	102	95	109	104
128	98	94	105	119
89	99	111	107	100

- Podajte nepristransko oceno povprečnega IQ v populaciji.

$$\bar{x} = \frac{1}{15}\Big(117 + 102 + \ldots + 100\Big) = 105,13$$

- Podajte nepristransko oceno za varianco populacije.

$$s^2 = \frac{1}{14}\Big((117 - 105,13)^2 + \ldots + (100 - 105,13)^2\Big) = 108.7$$

Primer. Iz populacije velikosti 1000 naključno izberemo 15 ljudi in jih prosimo, da rešijo IQ test. Dobimo nasledje rezultate

117	102	95	109	104
128	98	94	105	119
89	99	111	107	100

- Podajte nepristransko oceno povprečnega IQ v populaciji.

$$\bar{x} = \frac{1}{15}\Big(117 + 102 + \ldots + 100\Big) = 105,13$$

- Podajte nepristransko oceno za varianco populacije.

$$s^2 = \frac{1}{14}\Big((117 - 105,13)^2 + \ldots + (100 - 105,13)^2\Big) = 108.7$$

Primer. Iz populacije velikosti 1000 naključno izberemo 15 ljudi in jih prosimo, da rešijo IQ test. Dobimo nasledje rezultate

117	102	95	109	104
128	98	94	105	119
89	99	111	107	100

- Podajte nepristransko oceno povprečnega IQ v populaciji.

$$\bar{x} = \frac{1}{15}\Big(117 + 102 + \ldots + 100\Big) = 105,13$$

- Podajte nepristransko oceno za varianco populacije.

$$s^2 = \frac{1}{14}\Big((117 - 105,13)^2 + \ldots + (100 - 105,13)^2\Big) = 108.7$$

- Denimo, da je IQ v populaciji normalno porazdeljen z $\mu = 100$ in $\sigma = 15$. Kako tipičen je zgornji vzorec?

  Izračunajmo delež tistih vzorcev velikosti $n = 15$, pri katerih povprečje preseže $\bar{x} = 105,13$.

$$\begin{aligned} P(\bar{X} \geq 105,13) &= P\left(Z \geq \frac{105,13 - 100}{\frac{15}{\sqrt{15}}}\right) \\ &= P(Z \geq 1.32) = P(Z \leq -1.32) \\ &= 0.0934 \end{aligned}$$

  Torej pri približno 9.3% vzorcev velikosti $n = 15$ bo povprečje IQ preseglo 105,13.