Бизнис статистика

Оценување на параметри

Вовед

- Како што дефиниравме претходно, статистиката е наука која се занимава со прибирање, организација, групирање и анализа на податоци, како и изведување на заклучоци кои важат за целокупното множество податоци, иако набљудувањата се направени на само дел од податоците (односно на случајно избран примерок).

- Процесот на статистичко заклучување е процес во кој треба да се извлечат информации од примерокот кои ќе важат за целата популација.

- Постојат два вида на процедури за донесување заклучоци:
  - Оценување
  - Тестирање на хипотези

Вовед

- Целта на оценувањето е да се определат вредностите на непознатите параметри на обележјето врз основа на примерокот.
- Постојат два вида на оценувачи:
  - Точкасти оценувачи
  - Интервални оценувачи
- Точкастите оценувачи даваат оценка на вредноста на непознат параметар користејќи единечна вредност или точка.
- Интервалните оценувачи даваат оценка на вредноста на непознатиот параметар со користење на интервал.

Точкаст оценувач

- ***Точкаст оценувач*** на непознат параметар од кој зависи распределбата на обележјето е случајна променлива (статистика) која зависи само од примерокот.
- За дадена реализација на примерокот се добива конкретна вредност на точкастиот оценувач која се нарекува ***оценка*** на параметарот.
- Вредности на оценувачот (оценките) даваат апроксимација на непознатиот параметар.
- Ако се земе друг примерок за истото обележје, оценката на непознатиот параметар најверојатно ќе биде различна.
- Недостаток на точкастото оценување е тоа што, ако статистиката е случајна променлива со непрекината распределба, тогаш веројатноста дека некоја од пресметаните оценки (вредности на оценувачот) е еднаква на вистинскиот параметарот е 0.
- И покрај тоа, точкастото оценување е значаен прв чекор во решавање на статистичкиот проблем.

Точкасти оценувачи

Непознат параметар	Точкаст оценувач
Математичко очекување (μ)	$\bar{X}$
Пропорција/веројатност (p)	$\hat{P}$
Дисперзија (σ^2)	S^2
Разлика на мат. очекувања на две обележја ($\mu_1 - \mu_2$)	$\bar{X}_1 - \bar{X}_2$

Интервални оценувачи

- Вообичаен начин да се воведе точност на оценувачот е преку интервално оценување.

- Исто така, заради варирањето на вредностите на точкастите оценувачи за различни примероци, посоодветен за оценување на непознат параметар на распределбата на обележјето е интервалниот оценувач.

- Се формира интервал (базиран на примерокот) кој со одредена веројатност ја содржи точната, но непознатата вредност на параметарот. Овој интервал е познат како ***интервал на доверба*** (ИД)

- Постојат повеќе начини за определување на интервали на доверба, но ние ќе се задржиме на едноставен начин базиран на својствата на просекот и дисперзијата на случајниот примерок.

Интервали на доверба и ниво на доверба

- Нека θ е непознат параметар на распределбата на обележјето X и нека α е број помеѓу 0 и 1.

- Ако $P\{a < \theta < b\} = 1 - \alpha$, тогаш интервалот (a, b) се нарекува $(1 - \alpha)\cdot 100\%$ интервал на доверба за параметарот θ. Притоа, границите на интервалот се функции од примерокот.

- Вредноста $(1 - \alpha)$ се нарекува ***ниво на доверба*** на интервалот.

- Ако ја повторуваме постапката на земање примерок од популацијата и за секој од примероците го определуваме интервалот на доверба, тогаш $(1 - \alpha)100\%$ од интервалите пресметани на овој начин, ќе ја содржат вистинската вредност на параметарот θ.

Интерпретација на интервалите на доверба

- Важно е да се разбере интерпретацијата на интервалите на доверба.

- Имено, "довербата" се однесува на процедурата. Имаме процедура која креира интервали, така што $(1-\alpha)100\%$ од овие интервали на доверба ја содржат вистинската вредност на параметарот.

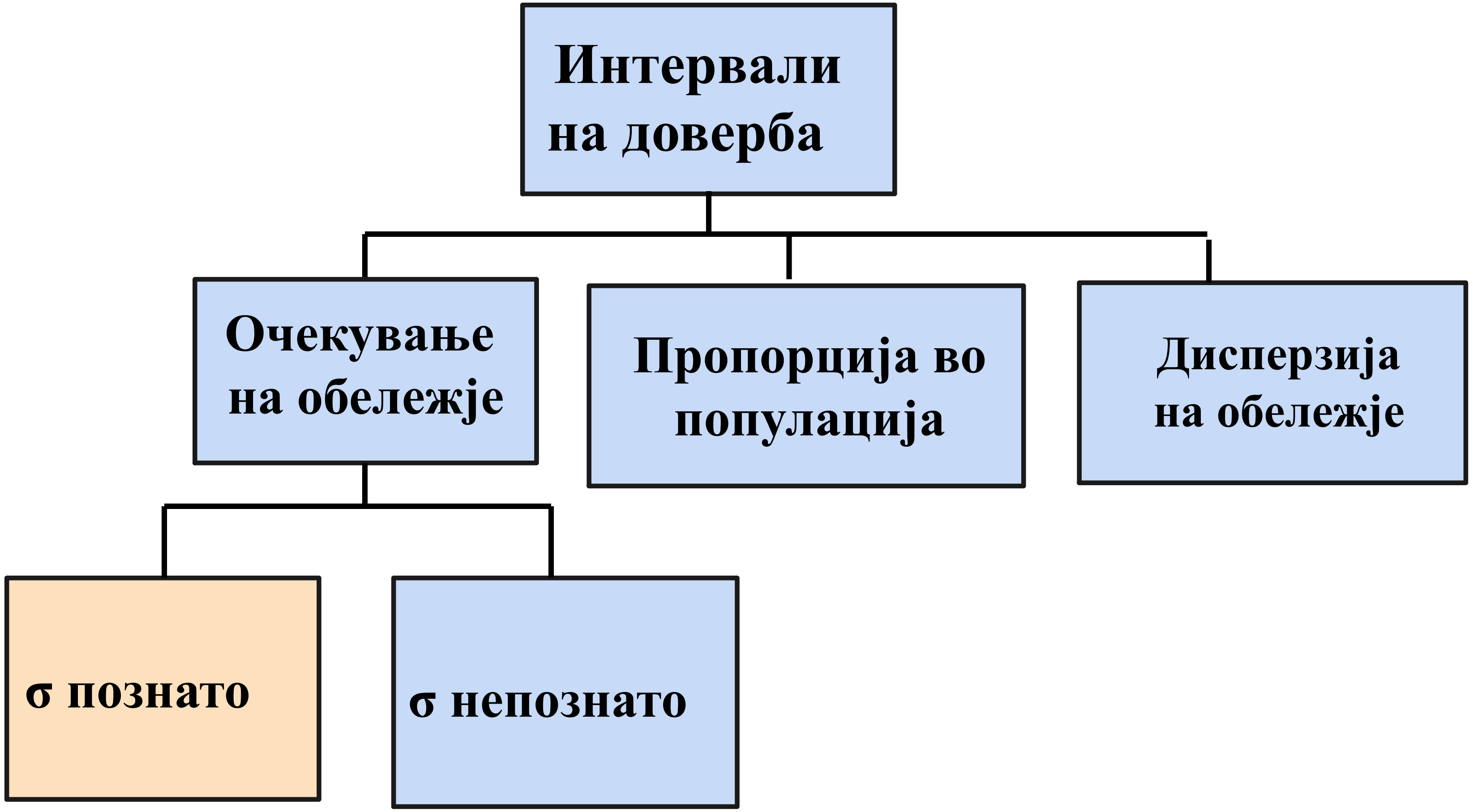
- При секоја дадена реализација на примерокот, интервалот на доверба или ја содржи или не ја содржи вистинската вредност на параметарот.

- Гаранцијата е само за просечното однесување при повторување на постапката.

Интерпретација на интервалите на доверба

- На пример, нека нивото на доверба е 95%, може да се запише и со $1 - \alpha = 0.95$
- Интерпретација на 95% интервал на доверба:
  - При повторување на постапката на земање на примерок и определување на границите на интервалот на доверба, 95% од добиените интервали на доверба (за различни примероци) ќе ја содржат вистинската, но непозната вредност на параметарот.
- Секој поединечен интервал или ќе ја содржи или нема да ја содржи точната вредност на параметарот.
- При определување на интервалите на доверба, се тргнува од статистика која е добар точкаст оценувач на непознатиот параметар и својствата на неговата распределба.

Интервали на доверба



Интервали на доверба за математичко очекување при познато σ^2

- Нека $\boldsymbol{X} = (X_1, \ldots, X_n)$ е случаен примерок од обележјето X за кое е исполнет еден од следните два услови:
  - Обележјето X има нормална распределба $N(\mu, \sigma^2)$, а обемот на примерокот е произволен, или
  - Обележјето X има било која распределба со математичко очекување μ и дисперзија σ^2 и $n > 30$.
- Притоа. $\mu \in R$ е непознат параметар, а параметарот σ^2 е познат.
- Знаеме дека случајната променлива

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$$

во првиот случај има $N(0,1)$ распределба, а во вториот случај има асимптотски $N(0,1)$ распределба, па оваа статистика ќе ја користиме за определување на интервалот на доверба.

Интервали на доверба за математичко очекување при познато σ^2

- За секое α помеѓу 0 и 1 може да најдеме броеви b и c такви што

$$P(b < Z < c) = 1 - \alpha \qquad (1)$$

- Ако замениме за Z добиваме

$$P\left(b < \frac{\bar{X} - \mu}{\sigma}\sqrt{n} < c\right) = 1 - \alpha$$

- Со преуредување на неравенството добиваме

$$P\left(\bar{X} - \frac{c\sigma}{\sqrt{n}} < \mu < \bar{X} - \frac{b\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

- Ако во претодното равенство, ставиме $D(\boldsymbol{X}) = \bar{X} - \frac{c\sigma}{\sqrt{n}}$ и $G(\boldsymbol{X}) = \bar{X} - \frac{b\sigma}{\sqrt{n}}$ велиме дека $(D(\boldsymbol{X}), G(\boldsymbol{X}))$ односно

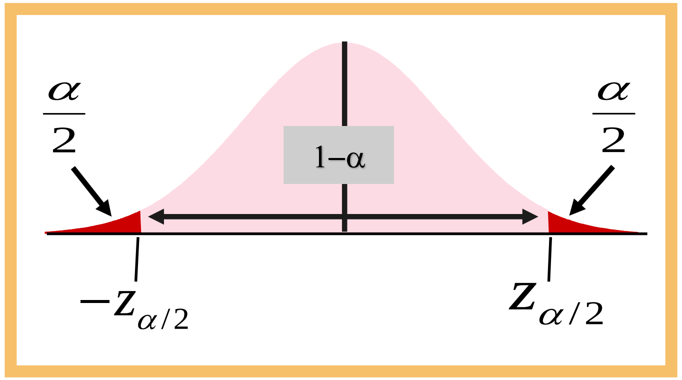
$$\left(\bar{X} - \frac{c\sigma}{\sqrt{n}}, \bar{X} - \frac{b\sigma}{\sqrt{n}}\right)$$

претставува **интервал на доверба за μ со ниво на доверба $(1-\alpha)100\%$.**

Интервали на доверба за математичко очекување при познато σ^2

- Да забележиме дека може да се најдат бесконечно многу парови (b, c) кои го задоволуваат условот (1).
- Меѓу сите нив секако најпогоден ќе биде паротот кој дава најкраток интервал на доверба. Во овој случај тоа не е тешко да се направи, бидејќи знаеме дека густината на распределба на $N(0,1)$ е симетрична околу 0 и го достигнува својот максимум во 0.

- Затоа, најкус интервал ќе добиеме ако земеме $(b, c) = (-z_{\alpha/2}, +z_{\alpha/2})$.
- Вредноста $z_{\alpha/2}$ се чита од таблицата за нормална распределба така што $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.



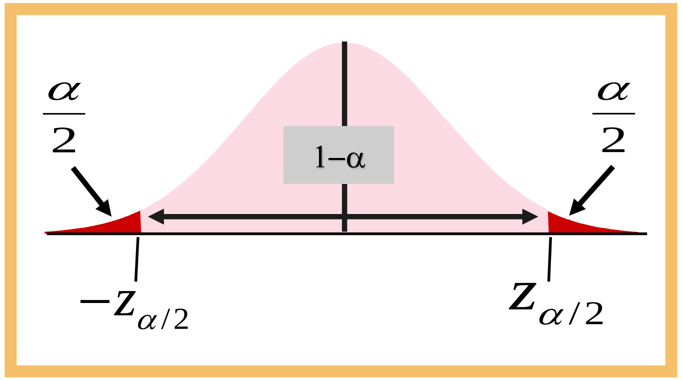
Интервали на доверба за математичко очекување при познато σ^2

- Значи, ако обележјето X има нормална распределба $N(\mu, \sigma^2)$ или $n > 30$ и X има било која распределба со математичко очекување μ и дисперзија σ^2, $(1-\alpha)100\%$ интервал на доверба за μ е:

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

- Вредноста $z_{\alpha/2}$ е таква што $\Phi(z_{\alpha/2}) = P(Z < z_{\alpha/2}) = 1 - \alpha/2$, и притоа

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$



Интервали на доверба за математичко очекување при познато σ^2

- Според тоа, интервалот на доверба за μ е

Критична вредност (Фактор на доверба)

Стандардна грешка

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Вредноста

$$e = ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

се нарекува и **грешка на примерок** или **маргина на грешка**

Прецизност на интервалите на доверба

- Ширината на интервалот на доверба е мерка за прецизноста на оценувањето.
- Да забележиме дека ширината на интервалот на доверба за математичкото очекување на обележјето е:

$$\Delta = \frac{2z_{\alpha/2}\sigma}{\sqrt{n}}$$

- На пример, ако бараме 95% ИД, тогаш $\alpha = 0.05$ и $\Phi(z_{0.025}) = 1 - 0.025 = 0.975$, па $z_{\alpha/2} = z_{0.025} = 1.96$ и ширината ќе биде $\Delta = 3.92\sigma/\sqrt{n}$.
- Ако пак бараме 99% ИД, $\alpha = 0.01$, $z_{0.005} = 2.58$, ширината ќе биде $\Delta = 5.16\sigma/\sqrt{n}$.
- Значи, ако се зголемува нивото на доверба при фиксно n, се намалува прецизноста, односно се зголемува ширината на интервалот.
- Доколку пак сакаме поголема прецизност за дадено ниво на доверба треба да го зголемиме обемот на примерокот.

Избор на големина на примерок

- Во повеќето бизнис истражувања, посебно е важно да се процени обемот на примерокот неопходен за да се постигнат зададените цели.

- На пример, ако голема корпорација прави истражување на пазарот, прашањето е дали да се земе примерок од 40 или од 4000 луѓе?

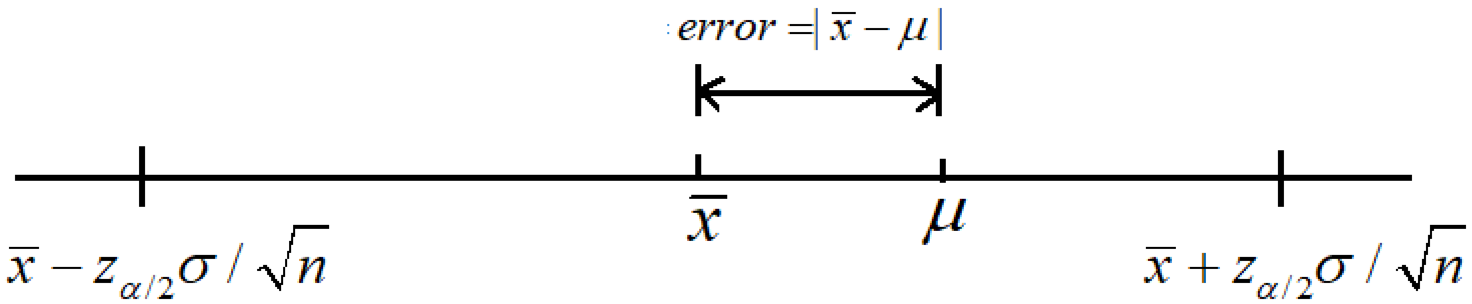
- Ова прашање е многу важно, затоа што најчесто ваквите истражувања се поврзани со одредени трошоци, па бизнис истражувачите не сакаат да земат поголем примерок од тоа што е неопходно.

Избор на големина на примерок

- Ширината на интервалот на доверба за математичкото очекување на обележјето е

$$\Delta = \frac{2z_{\alpha/2}\sigma}{\sqrt{n}}.$$

- Ова значи дека кога се користи $\bar{x}$ за проценка на μ, со доверба од $(1-\alpha)100\%$, грешката $error = |\bar{x} - \mu|$ е помала или еднаква од $z_{\alpha/2}\sigma/\sqrt{n}$:



Избор на големина на примерок

- Прашањето е како да се избере обемот на примерокот, така што грешката не надминува дадена вредност *E*?
- Од

$$\frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq E$$

наоѓаме

$$\frac{z_{\alpha/2}\sigma}{E} \leq \sqrt{n}, \qquad \text{т.е.} \qquad n \geq \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 .$$

Избор на големина на примерок

- Многу често во пракса, дисперзијата на обележјето не е позната, а пред да се земе примерок не е позната ниту дисперзијата на примерокот.
- Затоа, за прифатлива оценка на стандардната девијација во определување на потребната големината на примерокот (со формулата на претходниот слајд) се зема

$$\sigma \approx \tfrac{1}{4} \cdot (\text{распонот})$$

- Оваа вредност на σ се заменува во претходната формула.

Пример 1

Колкав обем на примерок треба да се земе за да се процени просечната месечната потрошувачка на мобилен интернет со ниво на доверба од 95% и притоа маргината на грешката што би ја толерирале да не е поголема од 10 MB. Од поранешните истажувања познато е дека стандардната девијација на потрошувачката е 30 MB.

Решение:

Сакаме грешката да не надминува вредност E = 10. За α = 0.05, $z_{\alpha/2} = z_{0.025} = 1.96$. Според формулата изведена во претходниот слајд обемот на примерокот се избира така што ќе важи:

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{E} \right)^2 .$$

За σ = 30 и E = 10, се добива

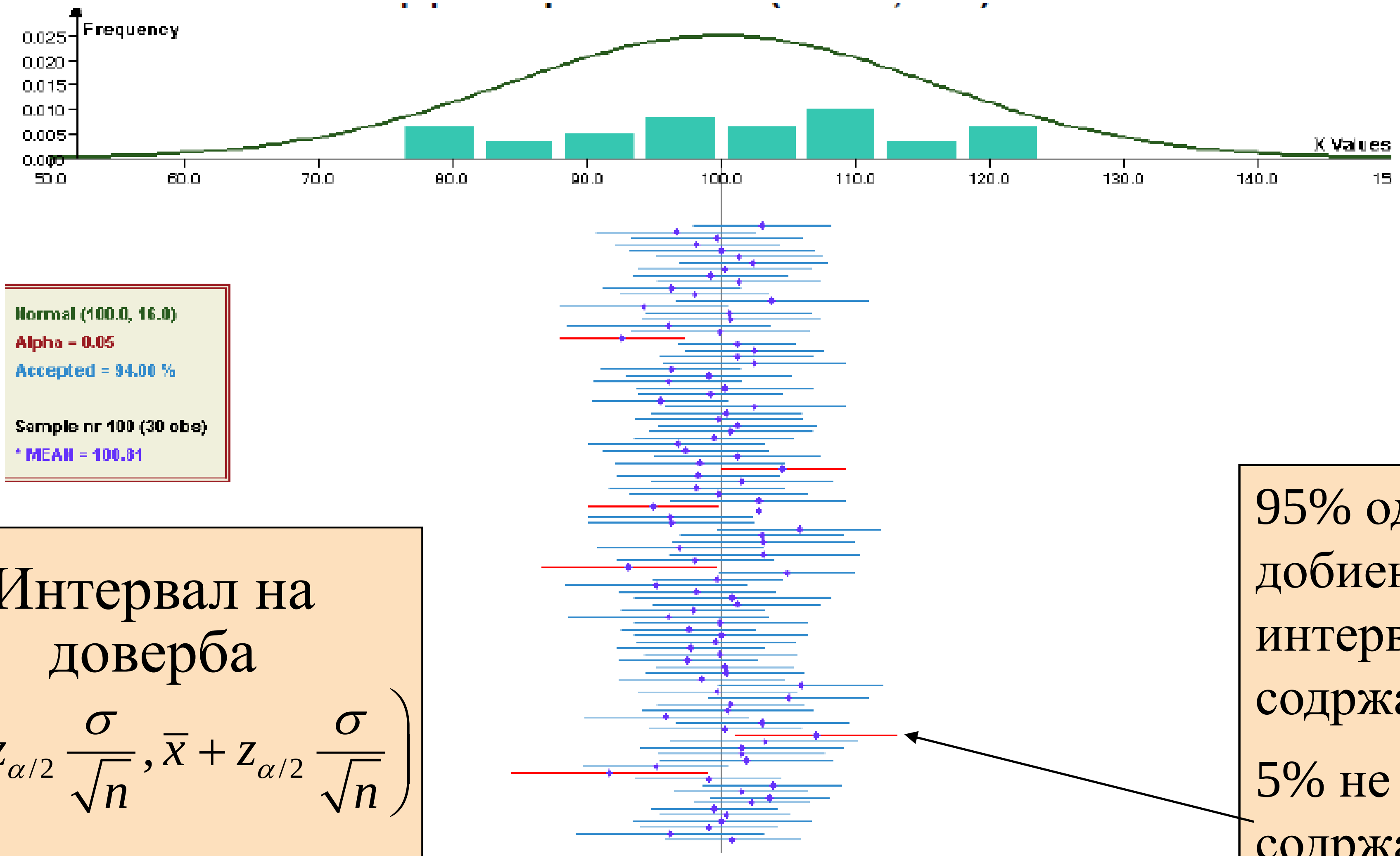
$$n \geq \frac{(1.96)^2 (30)^2}{10^2} = 34.57 \quad \Rightarrow \quad n \geq 35$$

Често користени нивоа на доверба

- Најчесто користени нивоа на доверба се 90%, 95%, и 99%.

Ниво на доверба	*Ниво на доверба,* $1 - \alpha$	$Z_{\alpha/2}$ *вредност*
80%	.80	1.28
90%	.90	1.65
95%	.95	1.96
98%	.98	2.33
99%	.99	2.58
99.8%	.998	3.08
99.9%	.999	3.27

Графички приказ на 95% интервали на доверба за математичко очекување на обележје со $N(100, 16)$



Интервал на доверба

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

95% од добиените интервали го содржат μ;

5% не го содржат

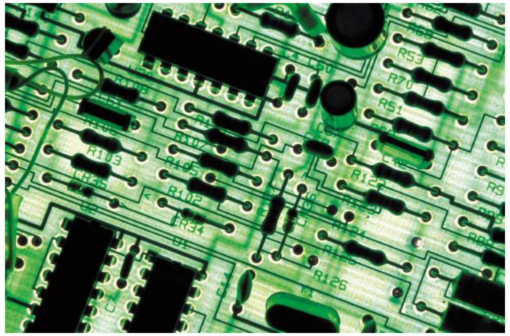
Пример 2

Просечниот отпор на примерок од 11 кола е 2.20 ома. Од претходни тестирања познато е дека отпорот во овие кола има нормална распределба со стандардната девијација 0.35 ома. Да се определи 95% интервал на доверба за очекуваниот отпор на колото.

Решение: $n = 11, \quad \bar{x} = 2.2, \quad \sigma = 0.35.$

Бараме 95% интервал на доверба, што значи $\alpha = 0.05$. Ја определуваме вредноста $z_{0.025}$ така што $\Phi(z_{0.025}) = 1 - 0.025 = 0.975$. Од таблица за нормална нормирана распределба читаме дека $z_{\alpha/2} = z_{0.025} = 1.96$. За границите на интервалот на доверба се добива

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 2.20 \pm 1.96 \frac{0.35}{\sqrt{11}} = 2.20 \pm 0.2068$$



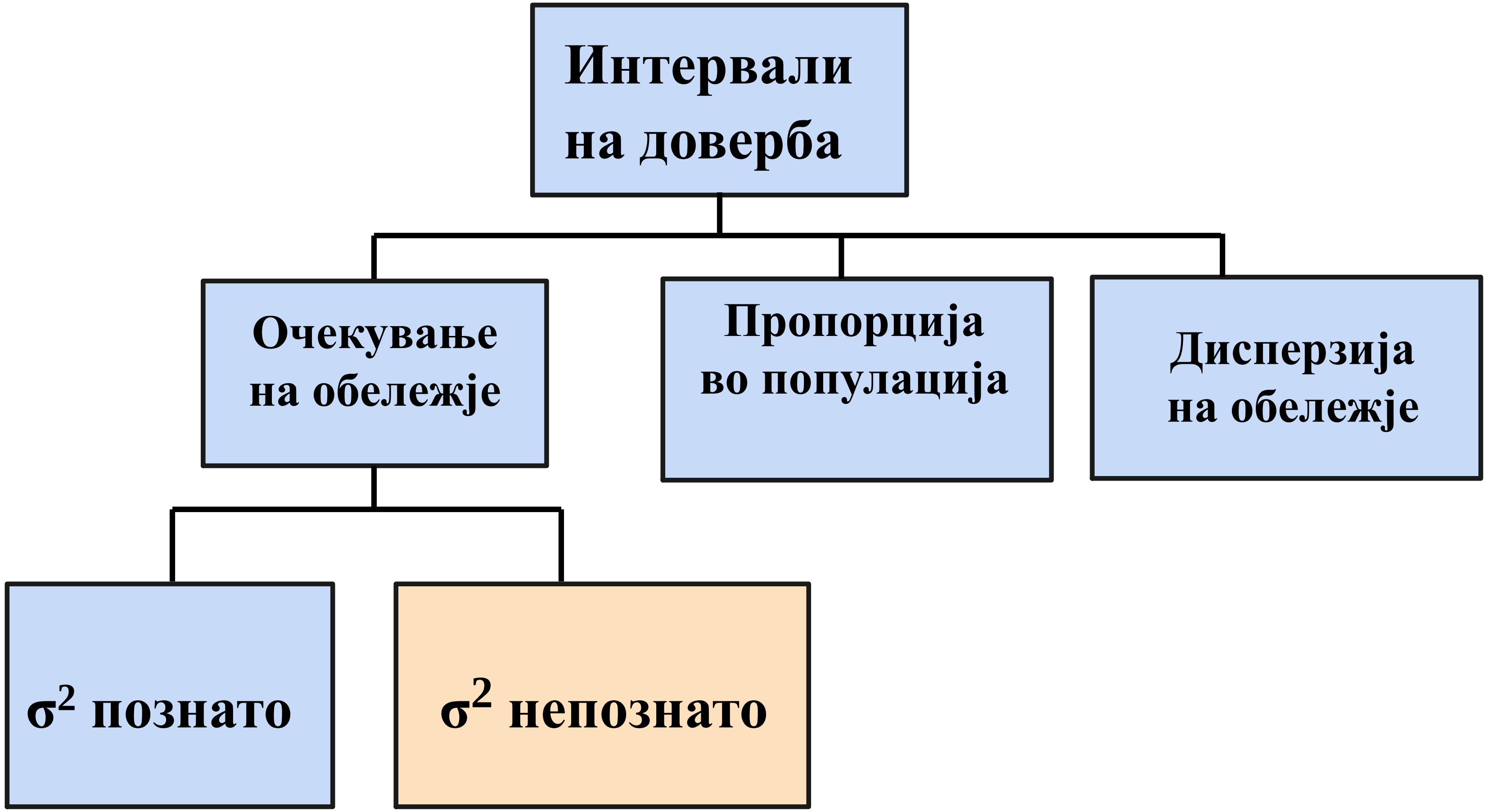
Значи 95% интервал на доверба за μ е (1.9932, 2.4068)

Интерпретација

- Со сигурност од 95% вистинскиот просечен отпор е меѓу 1.9932 и 2.4068 ома.

- Сепак, вистинскиот просек може да е или да не е во овој интервал. Ако се земат повеќе реализации на случајниот примерок и за секоја од нив се пресмета соодветниот интервал на доверба, 95% од интервалите добиени на овој начин ќе ја содржат вистинската вредност на математичкото очекување.

Интервали на доверба



Интервали на доверба за математичко очекување кај нормална распределба (σ непознато, мало *n*)

- Нека $(X_1, X_2, \ldots, X_n)$ е случаен примерок од обележје X со нормална распределба со непознати математичко очекување μ и дисперзија σ^2, и нека $(1-\alpha)$ е бараното ниво на доверба.
- Видовме дека статистиката

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

има t-распределба со $n - 1$ степени на слобода.

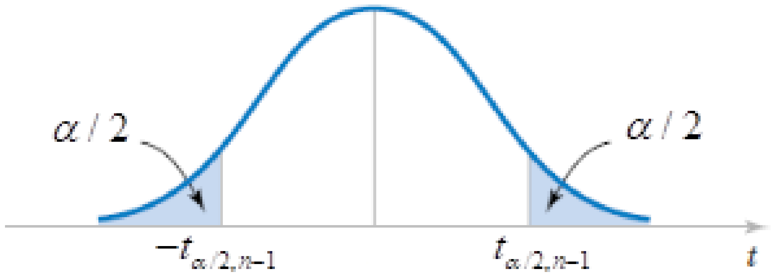
- Од симетричноста на t-распределбата и тоа што максимумот е во нулата се добива следното.

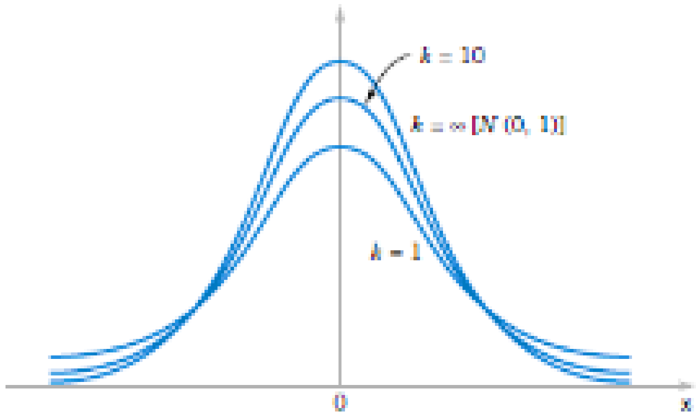
t - интервал на доверба за математичко очекување

$$P\left(-t_{\alpha/2,n-1} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2,n-1}\right) = 1-\alpha$$

Со средување на изразот добиваме

$$P\left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$





Овој интервал е познат како *t*-интервал на доверба и се користи кога обемот на примерокот $n < 30$, бидејќи кога бројот на степени расте *t*-распределбата тежи кон нормална нормирана распределба и за $n \geq 30$, апроксимацијата е точна скоро до 4-та децимала.

Интервали на доверба за математичко очекување, σ^2 непознато, голем примерок $n > 30$

- Со зголемување на степенот на слобода t - распределбата се доближува до нормална нормирана распределба.
- Затоа, ако дисперзијата на обележјето не е позната, но имаме доволно голем примерот ($n > 30$), во границите на интервалот на доверба за математичкото очекување се користат z – вредностите, наместо t – вредностите.
- Значи, ако параметарот σ^2 е непознат, **а обемот на примерокот е голем**, $(1 - \alpha)100\%$ интервал на доверба за математичкото очекување μ е:

$$\left(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

- Да нагласиме дека во случај на голем примерок ($n > 30$), овој интервал на доверба добива за произволна распределба, не само за нормална.

Пример 3

Случаен примерок со обем n = 25 за обележје X со нормална распределба, има просек $\bar{x}$ = 50 и стандардна девијација s = 8. Да се определи 95% интервал на доверба за математичкото очекување на X.

Решение: Имаме мал примерок за обележје со нормална распределба со непозната дисперзија. Затоа, ќе го користиме t интервалот на доверба.

Бидејќи, n = 25, степените на слобода t распределбата се $n - 1 = 24$, а $\alpha/2 = 0.05/2 = 0.025$

Од таблица за t распределба ја читаме вредноста $t_{24,\ 0.025} = 2.0639$.

$$\bar{x} - t_{n\text{-}1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n\text{-}1,\alpha/2}\frac{s}{\sqrt{n}}$$

$$50 - (2.0639)\frac{8}{\sqrt{25}} < \mu < 50 + (2.0639)\frac{8}{\sqrt{25}}$$

$$46.698 < \mu < 53.302$$

Интервалот на доверба е (46.698, 53.302)

Интервали на доверба за математичко очекување

- интервали на доверба за мат. очекување
  - сите распределби голем примерок ($n \geq 30$)
    - позната дисперзија — $\bar{X} \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$
    - непозната дисперзија — $\bar{X} \pm z_{\alpha/2} \dfrac{S}{\sqrt{n}}$
  - нормална распределба мал примерок
    - позната дисперзија — $\bar{X} \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$
    - непозната дисперзија — $\bar{X} \pm t_{\alpha/2,n-1} \dfrac{S}{\sqrt{n}}$

Пример 4

Да се конструира 95% интервал на доверба за математичкото очекување на обележје со нормална распределба врз база на следниот примерок:

2.5, 7.4, 8.0, 4.5, 7.4, 9.2

Мерниот инструмент гарантира стандардна девијација од $\sigma = 2.2$.

Решение:

Примерокот е со обем $n = 6$ и просек $\bar{x} = 6.5$.

За да добиеме ниво на доверба 95% $(1-\alpha = 0.95)$, потребно е $\alpha = 1 - 0.95 = 0.05$, т.е. $\alpha / 2 = 0.025$.

Од таблица читаме дека $z_{0.025}$ така што $\Phi(z_{0.025}) = 1 - 0.025 = 0.975$, $z_{0.025} = 1.96$.

Го добиваме следниот интервал на доверба

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 6.5 \pm (1.96)\frac{2.2}{\sqrt{6}} = 6.50 \pm 1.76 \text{ или } (4.74, 8.26)$$

Пример 5

Од претходно искуство познато е дека тежината на пастрмка одгледувана во еден рибник има нормална распределба со просек кој варира од сезона до сезона, но има фиксна стандардна девијација од 0.3. Доколку сакаме со 95% ниво на доверба да сме сигурни дека проценката на просечната тежина на пастрмката оваа сезона е точна во граници од $\pm 0.01\ kg$, колкав примерок треба да земеме?

Решение: 95% интервал на доверба за μ, базиран на примерок со обем n е:

$$\left(\bar{x} - 1.96 \cdot \sigma / \sqrt{n}, \bar{x} + 1.96 \cdot \sigma / \sqrt{n}\right)$$

Во овој случај сакаме ширината на интервалот да не надминува 0.02. Затоа, треба да го определиме n од неравенството

$$2 \cdot 1.96 \frac{\sigma}{\sqrt{n}} \leq 0.02$$

Добиваме

$$n \geq \left(1.96 \frac{\sigma}{0.01}\right)^2 = \left(\frac{1.96 \cdot 0.3}{0.01}\right)^2 = 3457.44 \Rightarrow n = 3458$$

Пример 6

За следниот примерок да се најде 90% интервал на доверба за математичкото очекување, ако се претпостави нормална распределба на обележјето.

0.46, 0.38, 0.31, 0.24, 0.20, 0.31, 0.34, 0.42, 0.09, 0.18, 0.46, 0.21

Решение:

$$n = 12$$

$$s = 0.1183$$

$$\bar{x} = 0.3$$

$$\alpha = 0.1, \quad t_{\alpha/2,n-1} = t_{0.05,11} = 1.796$$

90% интервал на доверба е

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 0.3 \pm 1.796 \cdot \frac{0.1183}{\sqrt{12}} = 0.3 \pm 0.0613 \quad \text{или} \quad (0.2387, 0.3613).$$

Пример 7

За да се утврди ефикасноста на еден сервер потребно е да се процени бројот на истовремени корисници. Според спроведени мерења, просечниот број на истовремени корисници измерени во 100 случајно избрани времиња е 37.7 со стандардна девијација 9.2. Да се конструира 90% интервал на доверба за очекуваниот број на истовремени корисници.

Решение:

$$n = 100,\ s = 9.2,\ \bar{x} = 37.7$$

$$n > 30 \text{ голем примерок}$$

$$\alpha = 0.1,\ \Phi(z_{\alpha/2}) = 1 - 0.05 = 0.95 \Rightarrow z_{\alpha/2} = 1.65$$

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 37.7 \pm z_{0.05} \frac{9.2}{10} = 37.7 \pm 1.5134 \quad \text{или} \quad (36.1866, 39.2134).$$

Пример 8

Сигнал кој има нормална распределба $N(\mu, 4)$ се пренесува од локација A до локација B. Сигналот е пратен 9 пати, при што се добиваат следните податоци:

5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5.

а) Да се конструира 95% интервал на доверба за μ.

б) Да се конструира 99% интервал на доверба за μ.

Решение: $\bar{x} = \frac{81}{9} = 9$

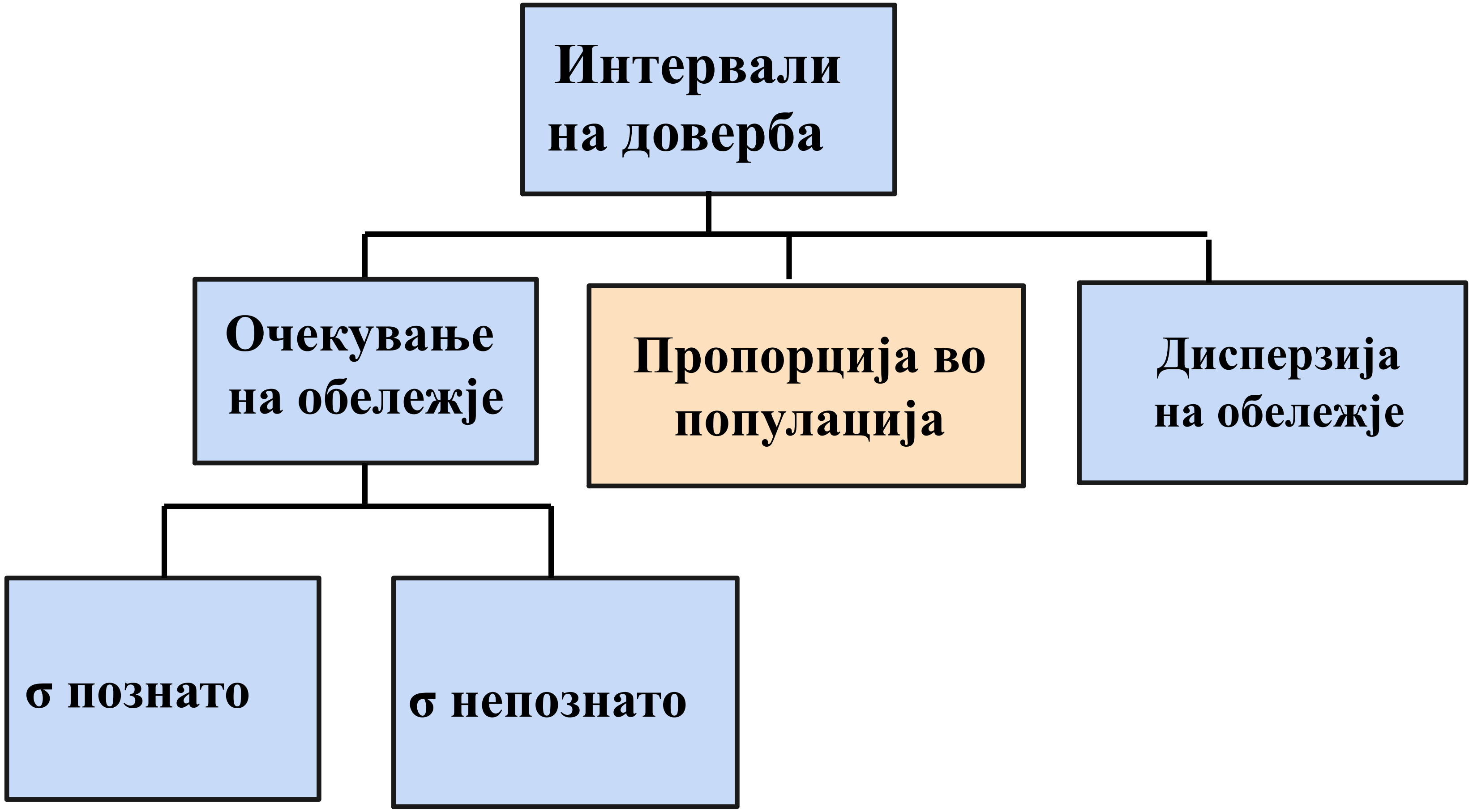
а) $\alpha/2 = 0.025, \;\; z_{0.025} = 1.96, \;\; \left(9 - 1.96 \cdot \frac{2}{\sqrt{9}}, 9 + 1.96 \cdot \frac{2}{\sqrt{9}}\right) = (7.69, 10.31)$

б) $\alpha/2 = 0.005, \;\; z_{0.005} = 2.58, \;\; \left(9 - 2.58 \cdot \frac{2}{\sqrt{9}}, 9 + 2.58 \cdot \frac{2}{\sqrt{9}}\right) = (7.28, 10.72)$

Може да забележиме дека 99% интервал на доверба е поширок од 95% интервал на доверба. Средината на двата интервали е $\bar{x} = 9$ ((10.31+7.69)/2 = (10.72+7.28)/2 = 9).

За кој било интервал на доверба за мат. очекување, средината на интервалот е просекот на примерокот, а со зголемување на нивото на доверба се зголемува ширината на интервалот.

Интервали на доверба



Интервал на доверба за веројатност на настан, односно пропорција во популација

- Нека p е веројатноста од популацијата да се избере единка која има одредена карактеристика, т.е. пропорција на единките со одредена карактеристика во целата популација. За определување на интервалниот оценувач може да се тргне од пропорцијата (статистиката) $\hat{P}$ на единките во примерокот со разгледуваната карактеристика.

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

каде X_i се случајни променливи со Бернулиева распределба. Тие примаат вредност 1, ако се избере единка со одредената карактеристика. Во останатите случаи, овие случајни променливи примаат вредност 0.

Интервал на доверба за веројатност на настан, односно пропорција во популација

- Нека $(X_1, \ldots, X_n)$ е случаен примерок од Бернулиева распределба, односно за статистичко обележје X - индикатор на настан со непозната веројатност на појавување p.
- Видовме дека за големо n, статистика

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

има приближно нормална распределба со математичко очекување p и дисперзија $p(1-p)/n$.

- Според тоа статистиката

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

има $N(0,1)$ распределба.

Интервал на доверба за веројатност на настан (голем примерок)

- Од тука, добиваме

$$P\left(-z_{\alpha/2} < \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

- Забележуваме дека границите на интервалот зависат од p, параметарот кој го оценуваме,.
- Но, задоволителни резултати се добиваат ако во границите параметарот p се замени со оценувачот $\hat{P}$.
- Според тоа за $(1-\alpha)100\%$ интервал на доверба за непознатата пропорција p во популацијата се добива

$$\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right)$$

Пример 9

Од 1000 новороденчиња родени во една болница, 517 биле машки. Да се определи 90% интервал на доверба за процентот на машки новороденчиња.

Решение:

$$n = 1000, \quad \alpha = 0.1, \quad z_{\alpha/2} = z_{0.05} = 1.65,$$

$$\hat{p} = \frac{517}{1000} = 0.517,$$

Бараниот интервал на доверба е:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) =$$

$$= \left(0.517 - 1.65\sqrt{\frac{0.517(1-0.517)}{1000}}, 0.517 + 1.65\sqrt{\frac{0.517(1-0.517)}{1000}} \right) = (0.491, 0.543)$$

Пример 10

Во една статија New York Times информирал дека во анкетатата што ја направиле меѓу граѓаните, 52% од испитаниците би гласале за Хилари Клинтон, со маргинална грешка од $\pm 4\%$. Ако се знае дека весникот пресметувал 95% интервал на доверба, може ли да се определи колку граѓани биле анкетирани?

Решение: Бидејќи $z_{\alpha/2} = z_{0.025} = 1.96$, 95% ИД за p - процентот на популацијата кој ќе гласа за Х. Клинтон е даден со

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = \left(0.52 - 1.96\sqrt{\frac{0.52 \cdot 0.48}{n}}, 0.52 + 1.96\sqrt{\frac{0.52 \cdot 0.48}{n}}\right)$$

каде n е големината на примерокот.

Бидејќи маргиналната грешка е ± 4 проценти, ширината на интервалот е 2·0.04, т.е.,

$$2 \cdot 1.96\sqrt{\frac{0.52 \cdot 0.48}{n}} = 2 \cdot 0.04$$

Од каде $n = 1.96^2 \cdot \frac{0.52 \cdot 0.48}{0.04^2} = 599.29$. Заклучуваме дека приближно 599 граѓани биле испитани и 52% од нив би гласале за Хилари Клинтон.

Пример 11

Во примерок од 100 луѓе 25 се левораки. Да се определи 95% интервал на доверба за вистинската пропорција на левораки во популацијата.

Решение: $n = 100, \quad \alpha = 0.05, \quad z_{\alpha/2} = z_{0.025} = 1.96,$

$$\hat{p} = \frac{25}{100} = 0.25,$$

Бараниот интервал на доверба е:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) =$$

$$= \left(0.25 - 1.96\sqrt{\frac{0.25(1-0.25)}{100}}, 0.25 + 1.96\sqrt{\frac{0.25(1-0.25)}{100}} \right) = (0.1651, 0.3349)$$

Ние сме 95% уверени дека вистинскиот процент на левораки луѓе во популацијата е помеѓу 16.51% и 33.49%. Иако, интервалот може да ја содржи или да не ја содржи вистинската пропорција, 95% од сите интервали добиени на овој начин за различни примероци со обем 100 ќе ја содржат вистинската пропорција.

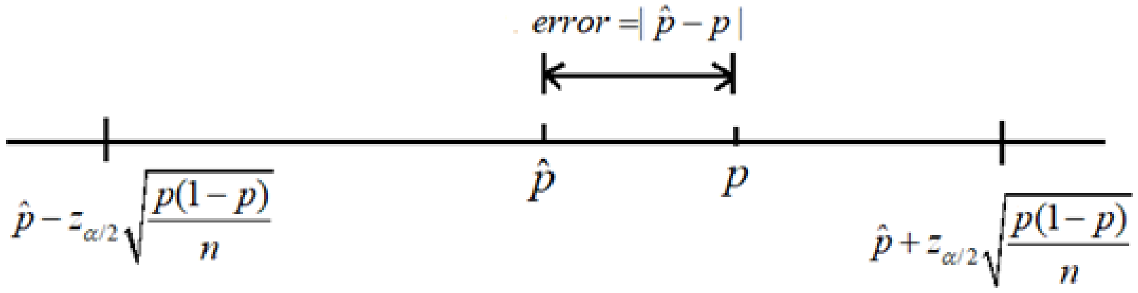
Избор на обем на примерок за оценување на p

- Грешката на оценување на непознатата пропорција p со точкастата оценка $\hat{p}$ е

$$error = |\hat{p} - p|$$

- За ниво на значајност $1 - \alpha$, имаме дека

$$|\hat{p} - p| < z_{\alpha/2}\sqrt{p(1-p)/n}.$$



- Ако сакаме грешката да не надминува дадена вредност E треба

$$z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le E.$$

- Ако решиме по n добиваме

$$n \ge \frac{z_{\alpha/2}^2 p(1-p)}{E^2}$$

што зависи од p.

Избор на обем на примерок

- Многу ретко однапред имаме сознание за вредноста на параметарот p и затоа во формулата за обемот на примерокот треба да се избере онаа вредност на p, за која n би било најголемо. Најголема вредност за n се добива за p = 0.5, т.е., за дадена (дозволена) грешка E и ниво на доверба $1-\alpha$, за обемот на примерок добиваме:

$$n \geq \frac{z_{\alpha/2}^2 (0.5)^2}{E^2} = \frac{z_{\alpha/2}^2}{4E^2} = \frac{z_{\alpha/2}^2}{(2E)^2}$$

Пример 12

- Колкав е најмалиот обем на примерок од избирачи кои треба да се интервјуираат по однос на тоа дали се за или против одреден кандидат, ако сакаме маргината на грешка да не е поголема од 1% со ниво на доверба од 95%.

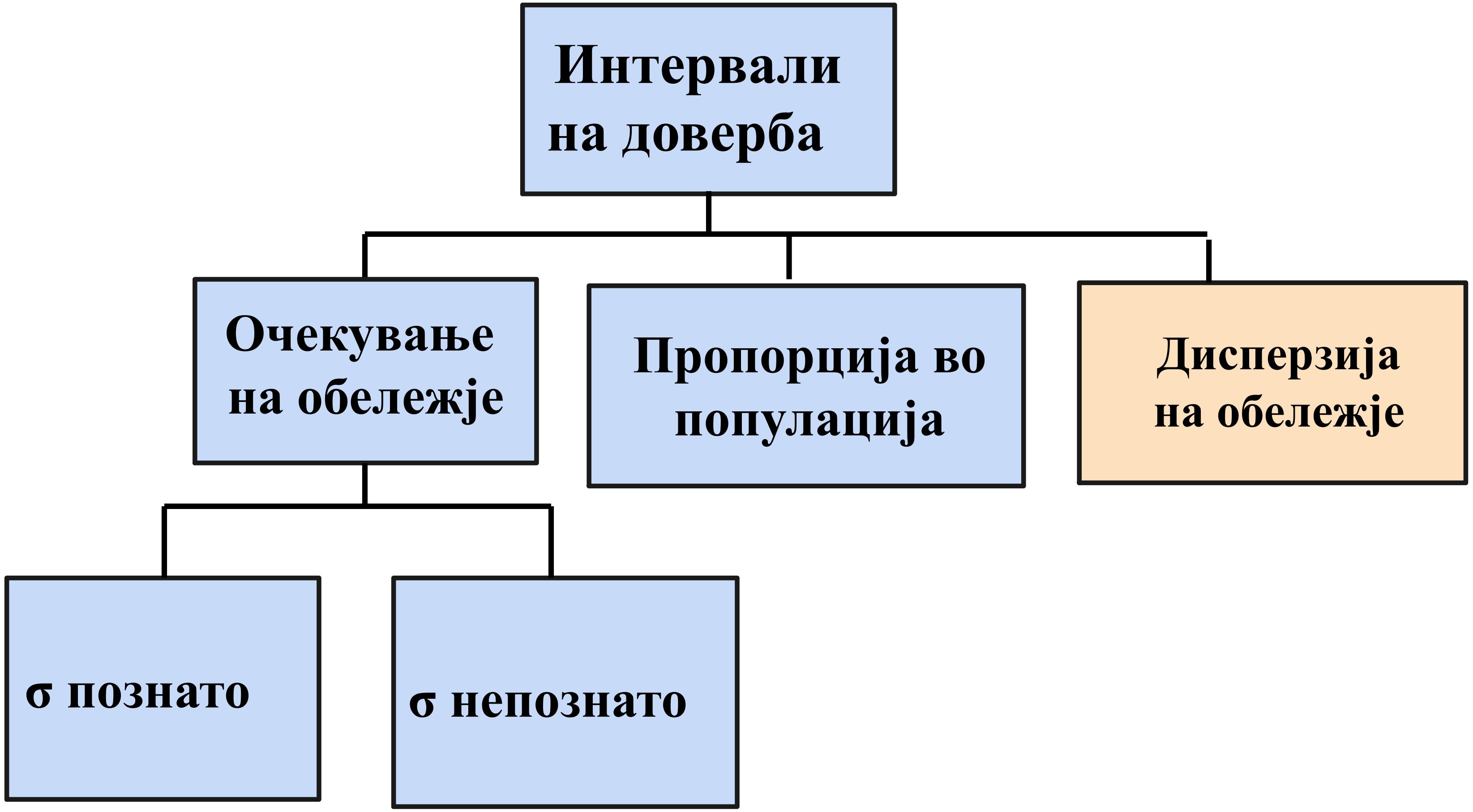
- **Решение**:

$$E = 0.01$$

95% ниво на доверба $\Rightarrow z = 1.96$

$$n \geq \frac{z_{\alpha/2}^2}{(2E)^2} = \frac{(1.96)^2}{(2 \cdot 0.01)^2} = 9604$$

Интервали на доверба



Интервали на доверба за дисперзијата, σ^2

- Дисперзијата е значаен индикатор за вкупниот квалитет на стандардизирани производи и услуги, а менаџерите се трудат да ги подобрат процесите со намалување на дисперзијата.
- Дисперзијата е мерка и за финансискиот ризик. Дисперзијата на ратата на повраток на сретства при одредени инвестиции овозможува да се бараат алтернативи за финансиски и капитални инвестиции.
- Варирањето е реалност во глобалните пазари, продуктивноста, платите, животните трошоци во различни региони и земји.
- Затоа, во бизнис истражувањето, дисперзијата на обележјето е многу битна.
- Статистиката S^2 (дисперзија на примерокот) е добар оценувач за σ^2 (дисперзијата на обележјето) и затоа за конструкција на $(1-\alpha)100\%$ интервал на доверба за σ^2, поаѓаме од распределбата на статистиката S^2.

Интервали на доверба за дисперзијата на обележје со нормална распределба

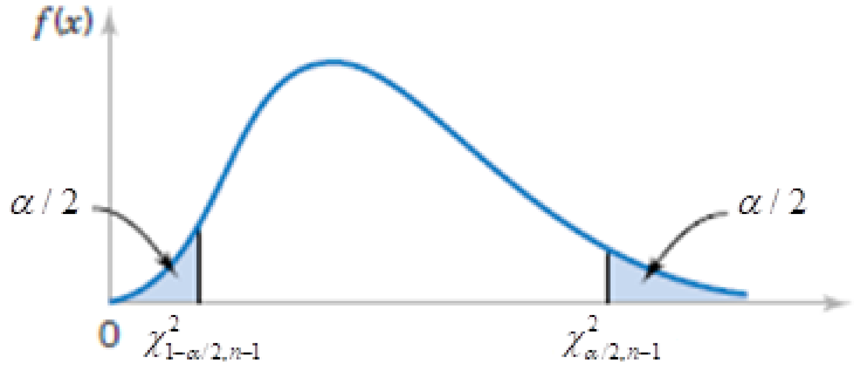
- Видовме дека ако обележјето има нормална $N(\mu, \sigma^2)$ распределба, тогаш статистиката

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

има χ^2 -распределба (Хи-квадрат распределба) со $n - 1$ степени на слобода.

- Според тоа

$$P\left(\chi^2_{1-\alpha/2,n-1} < \chi^2 < \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$



- Со замена на χ^2 добиваме:

$$P\left(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

Пример 13

Дисперзијата на примерок со обем 8 е $s^2 = 0.0022$. Да се определи 90% интервал на доверба за дисперзијата на обележјето кое има нормална распределба.

Решение:

$$s^2 = 0.0022,\ n = 8,\ \text{с.с.} = n-1 = 7,\ \alpha = 0.10$$

За интервалот на доверба се добива:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

Од таблица ги читаме вредностите:

$$\chi^2_{\frac{\alpha}{2},n-1} = \chi^2_{\frac{0.1}{2},7} = \chi^2_{0.05,7} = 14.0671$$

$$\chi^2_{1-\frac{\alpha}{2},n-1} = \chi^2_{1-\frac{0.1}{2},7} = \chi^2_{0.95,7} = 2.16735$$

$$\frac{(8-1)0.0022}{14.0671} < \sigma^2 < \frac{(8-1)0.0022}{2.16735}$$

$$0.001101 < \sigma^2 < 0.007146$$

Пример 14

Од обележје со нормална распределба земен е примерок со обем n = 27. Дисперзијата на примерокот е s^2 = 9.73. Да се определи 95% интервал на доверба за дисперзијата.

Решение:

$$n = 27$$

$$s^2 = 9.73$$

$$\alpha = 0.05$$

$$\chi^2_{\alpha/2,n-1} = \chi^2_{0.025,26} = 41.92, \quad \chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,26} = 13.84$$

Бараниот интервал на доверба е:

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right) = \left(\frac{(27-1)\cdot 9.73}{41.92}, \frac{(27-1)\cdot 9.73}{13.84}\right) = (6.035, 18.279)$$

Пример 15.

Заводот за статистика објавил дека просечната плата за 1 час на работниците во текстилната индустрија во Македонија е 48.3 денари. Синдикатот сака да знае колку е конзизтентен овој резултат. Затоа на случаен начин е избран примерок од 25 текстилни работници и добиено е дека стандардната девијација на платата во примерокот е 3.36 денари. Под претпоставка дека платата исплатена за 1 час има нормална распределба, да се определи 95% интервал на доверба за дисперзијата (варирањето) на просечната плата за еден час на вработените во текстилната индустрија.

Решение:

$$s^2 = 11.29, \quad n = 25, \quad c.c. = n-1 = 24, \quad \alpha = 0.05$$

$$\chi^2_{\frac{\alpha}{2},n-1} = \chi^2_{\frac{0.05}{2},24} = \chi^2_{0.025,24} = 39.3641$$

$$\chi^2_{1-\frac{\alpha}{2},n-1} = \chi^2_{1-\frac{0.05}{2},24} = \chi^2_{0.975,24} = 12.4011$$

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}$$

$$\frac{(25-1)(11.29)}{39.3641} < \sigma^2 < \frac{(25-1)(11.29)}{12.4011}$$

$$6.88 < \sigma^2 < 21.85$$

Интервали на доверба - заклучок

- На крајот, да потенцираме дека ако $(D(\mathbf{X}),G(\mathbf{X}))$ е $(1-\alpha)100\%$ интервал на доверба за непознатиот параметар θ на распределбата на обележјето моделирано со случајната променлива X, тогаш доколку земеме повеќе реализации на случајниот примерок $\boldsymbol{X} = (X_1, \dots, X_n)$, релативната честота на оние реализации на случајниот интервал $(D(\mathbf{X}),G(\mathbf{X}))$ кои ја содржат вистинската вредност на непознатиот параметар θ ќе се стабилизира околу бројот $1-\alpha$.
- При конкретна реализација $\boldsymbol{x} = (x_1, \dots, x_n)$ на случајниот примерокот $\boldsymbol{X}$ се добиваат конкретни вредности $d(\boldsymbol{x}) = d$ и $g(\boldsymbol{x}) = g$ и тогаш тврдењето $P(d < \theta < g) = 1-\alpha$, веќе не е точно тврдење бидејќи (d, g) веќе не е случаен интервал.