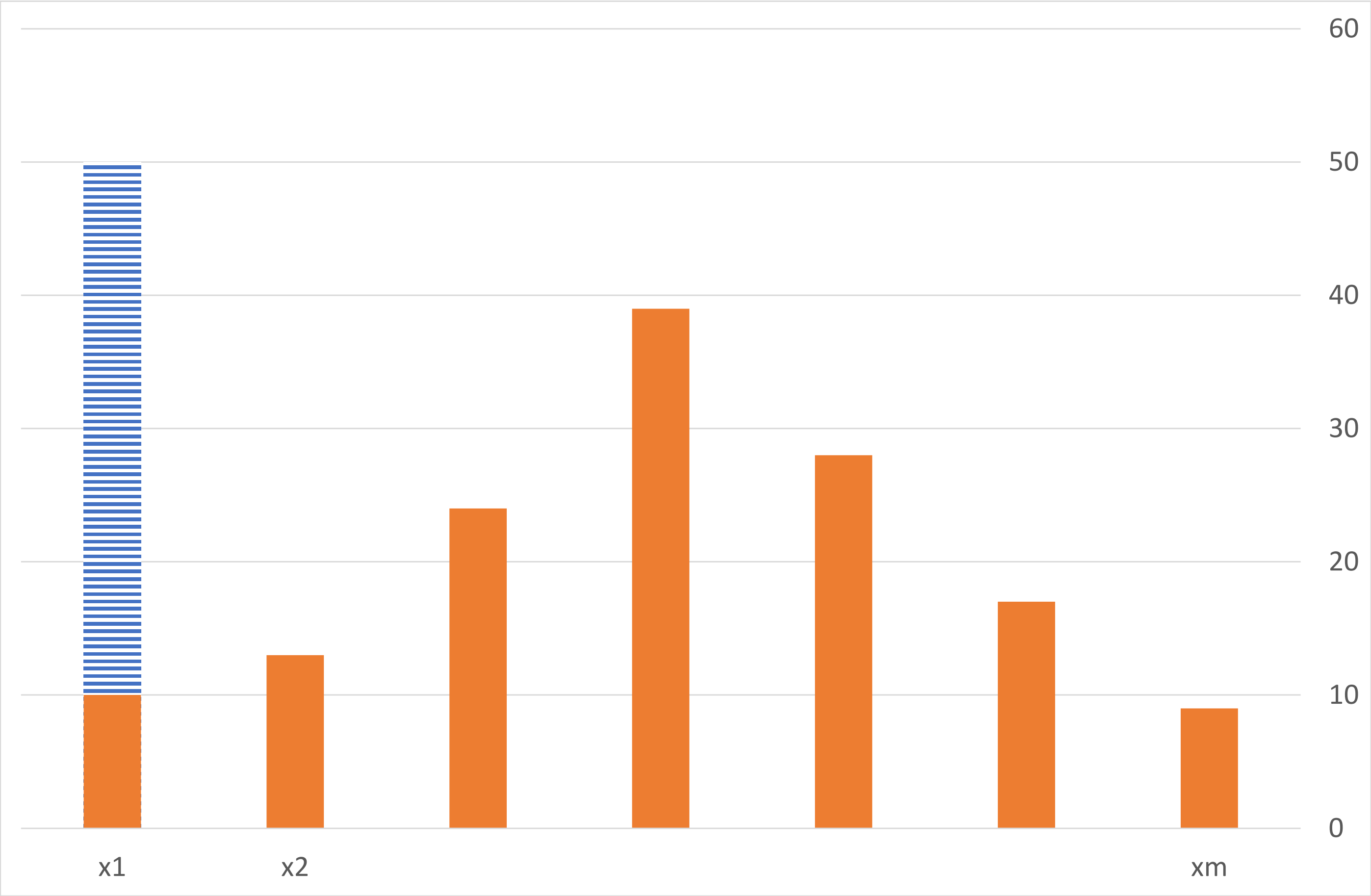
Статистически анализ на данни от непредставителни онлайн изследвания

Калоян Хараламииев

Обща постановка

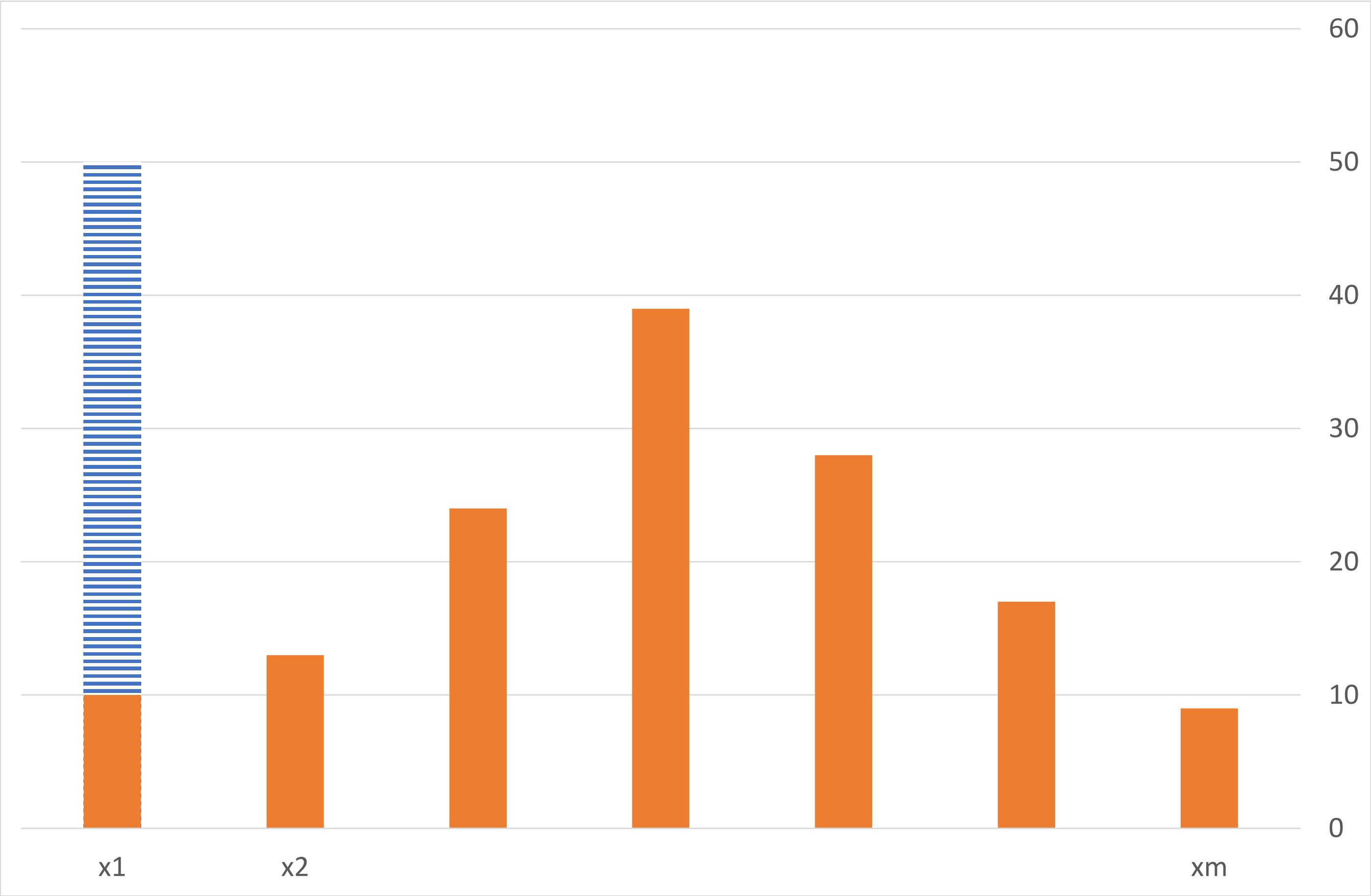


$$\sum f = n$$

$$\sum \hat{f} = N$$

Остават $N - n$ единици извън извадката.

Обща постановка

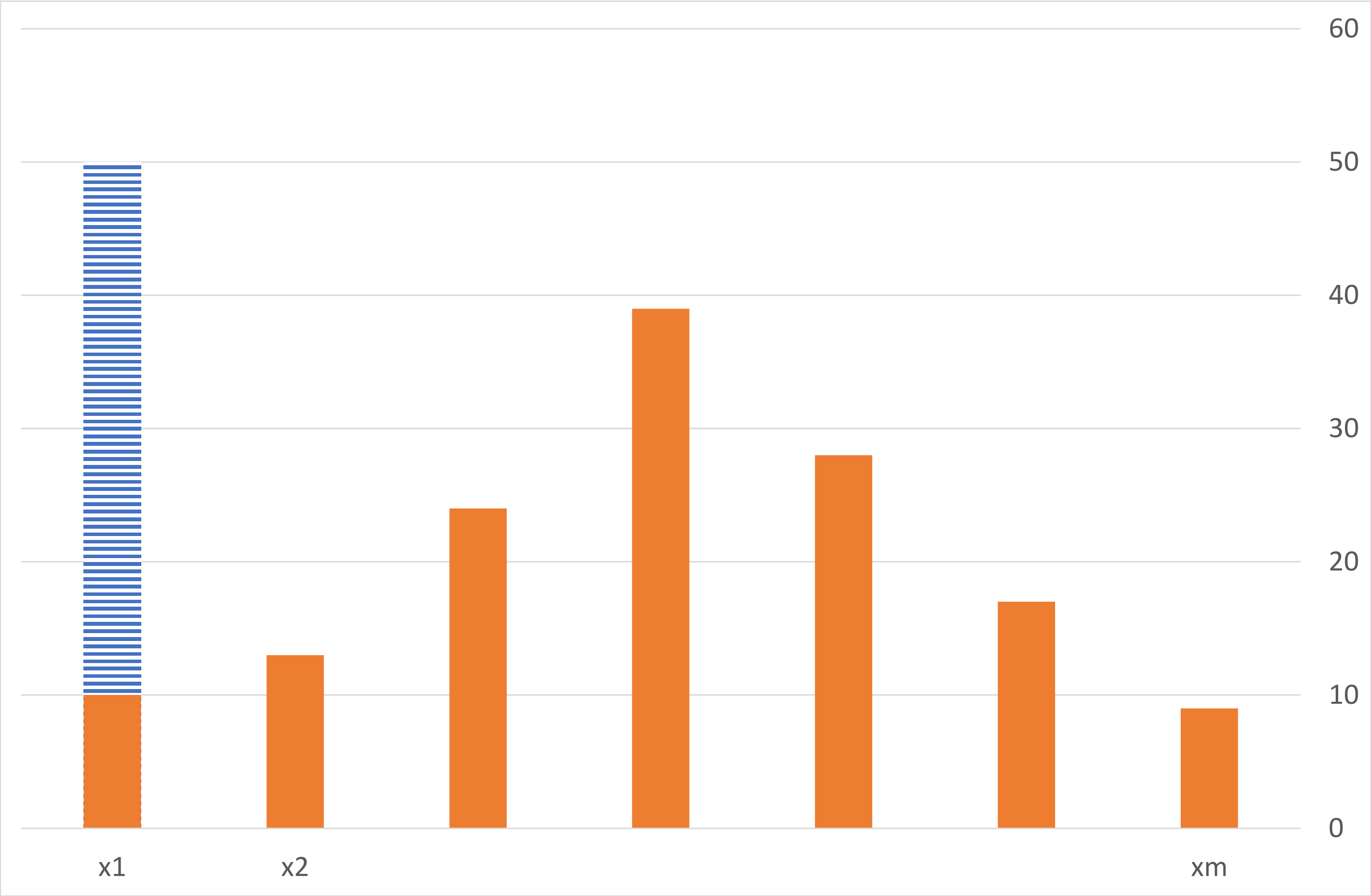


Оставащите $N-n$ единици могат да се разпределят в наличните m групи по $\tilde{C}_m^{N-n} = C_{N-n+m-1}^{N-n} = C_{N-n+m-1}^{m-1}$ начина.

Ако в една от групите има k допълнителни единици от оставащите $N-n$ единици ($0 \le k \le N-n$), тогава оставащите $N-n-k$ единици могат да се разпределят в оставащите $m-1$ групи по $\tilde{C}_{m-1}^{N-n-k} = C_{N-n-k+m-2}^{N-n-k} = C_{N-n-k+m-2}^{m-2}$ начина.

Тогава $P\left(\pi = \frac{f+k}{N}\right) = \frac{C_{N-n-k+m-2}^{m-2}}{C_{N-n+m-1}^{m-1}} = \frac{C_{N(1-\pi)-n(1-p)+m-2}^{m-2}}{C_{N-n+m-1}^{m-1}}$, където $p = \frac{f}{n}$ е относителният дял в извадката.

Обща постановка



Тогава плътността на разпределението е: $f(x) = P(\pi = x) = \frac{C_{N(1-x)-n(1-p)+m-2}^{m-2}}{C_{N-n+m-1}^{m-1}}$

Може да се докаже, че функцията на разпределението е: $F(x) = P(\pi \leq x) = 1 - \frac{C_{N(1-x)-n(1-p)+m-1}^{m-1}}{C_{N-n+m-1}^{m-1}}$

	Безвъзвратен подбор	Възвратен подбор
Малка генерална съвкупност	$\frac{f}{N} \leq \pi \leq \frac{f+N-n}{N}$ $f(x) = \frac{C_{N(1-x)-n(1-p)+m-2}^{m-2}}{C_{N-n+m-1}^{m-1}}$ $F(x) = 1 - \frac{C_{N(1-x)-n(1-p)+m-1}^{m-1}}{C_{N-n+m-1}^{m-1}}$	$\frac{1}{N} \leq \pi \leq \frac{1+N-m}{N}$ $f(x) = \frac{C_{N(1-x)-1}^{m-2}}{C_{N-1}^{m-1}}$ $F(x) = 1 - \frac{C_{N(1-x)}^{m-1}}{C_{N-1}^{m-1}}$
$\boldsymbol{N \to \infty; \frac{n}{N} \neq 0}$	$p\frac{n}{N} \leq \pi \leq p\frac{n}{N} + 1 - \frac{n}{N}$ $f(x) = \frac{m-1}{\left(1-\frac{n}{N}\right)^{m-1}}\left[1 - x - \frac{n}{N}(1-p)\right]^{m-2}$ $F(x) = 1 - \left[\frac{1 - x - \frac{n}{N}(1-p)}{1-\frac{n}{N}}\right]^{m-1}$	$0 \leq \pi \leq 1$ $f(x) = (m-1)(1-x)^{m-2}$ $F(x) = 1 - (1-x)^{m-1}$
$\boldsymbol{N \to \infty; \frac{n}{N} \to 0}$	$0 \leq \pi \leq 1$ $f(x) = (m-1)(1-x)^{m-2}$ $F(x) = 1 - (1-x)^{m-1}$	$0 \leq \pi \leq 1$ $f(x) = (m-1)(1-x)^{m-2}$ $F(x) = 1 - (1-x)^{m-1}$

	Безвъзвратен подбор	**Възвратен подбор**
Малка генерална съвкупност	$\frac{f}{N} \leq \pi \leq \frac{f+N-n}{N}$ $f(x) = \frac{C_{N(1-x)-n(1-p)+m-2}^{m-2}}{C_{N-n+m-1}^{m-1}}$ $F(x) = 1 - \frac{C_{N(1-x)-n(1-p)+m-1}^{m-1}}{C_{N-n+m-1}^{m-1}}$	$\frac{1}{N} \leq \pi \leq \frac{1+N-m}{N}$ $f(x) = \frac{C_{N(1-x)-1}^{m-2}}{C_{N-1}^{m-1}}$ $F(x) = 1 - \frac{C_{N(1-x)}^{m-1}}{C_{N-1}^{m-1}}$
$N \to \infty; \frac{n}{N} \neq 0$	$p\frac{n}{N} \leq \pi \leq p\frac{n}{N} + 1 - \frac{n}{N}$ $f(x) = \frac{m-1}{\left(1-\frac{n}{N}\right)^{m-1}}\left[1 - x - \frac{n}{N}(1-p)\right]^{m-2}$ $F(x) = 1 - \left[\frac{1 - x - \frac{n}{N}(1-p)}{1-\frac{n}{N}}\right]^{m-1}$	$0 \leq \pi \leq 1$ $f(x) = (m-1)(1-x)^{m-2}$ $F(x) = 1-(1-x)^{m-1}$
$N \to \infty; \frac{n}{N} \to 0$	$0 \leq \pi \leq 1$ $f(x) = (m-1)(1-x)^{m-2}$ $F(x) = 1-(1-x)^{m-1}$	$0 \leq \pi \leq 1$ $f(x) = (m-1)(1-x)^{m-2}$ $F(x) = 1-(1-x)^{m-1}$

Проверка на хипотези и доверителни интервали

$$P(\pi \leq a) = F(a)$$

$$P(\pi > a) = 1 - F(a)$$

$$P(a < \pi \leq b) = F(b) - F(a)$$

Най-тесен доверителен интервал

$$b = F^{-1}[F(b)]$$

$$a = F^{-1}[F(a)]$$

$$b - a = min \rightarrow (b - a)' = 0$$

Най-тесен доверителен интервал

$$(b-a)' = \{F^{-1}[F(b)] - F^{-1}[F(a)]\}'$$
$$= \{F^{-1}[F(b)]\}' - \{F^{-1}[F(a)]\}' = \frac{1}{F'(b)} - \frac{1}{F'(a)} = \frac{1}{f(b)} - \frac{1}{f(a)}$$
$$= \frac{f(a) - f(b)}{f(a)f(b)}$$

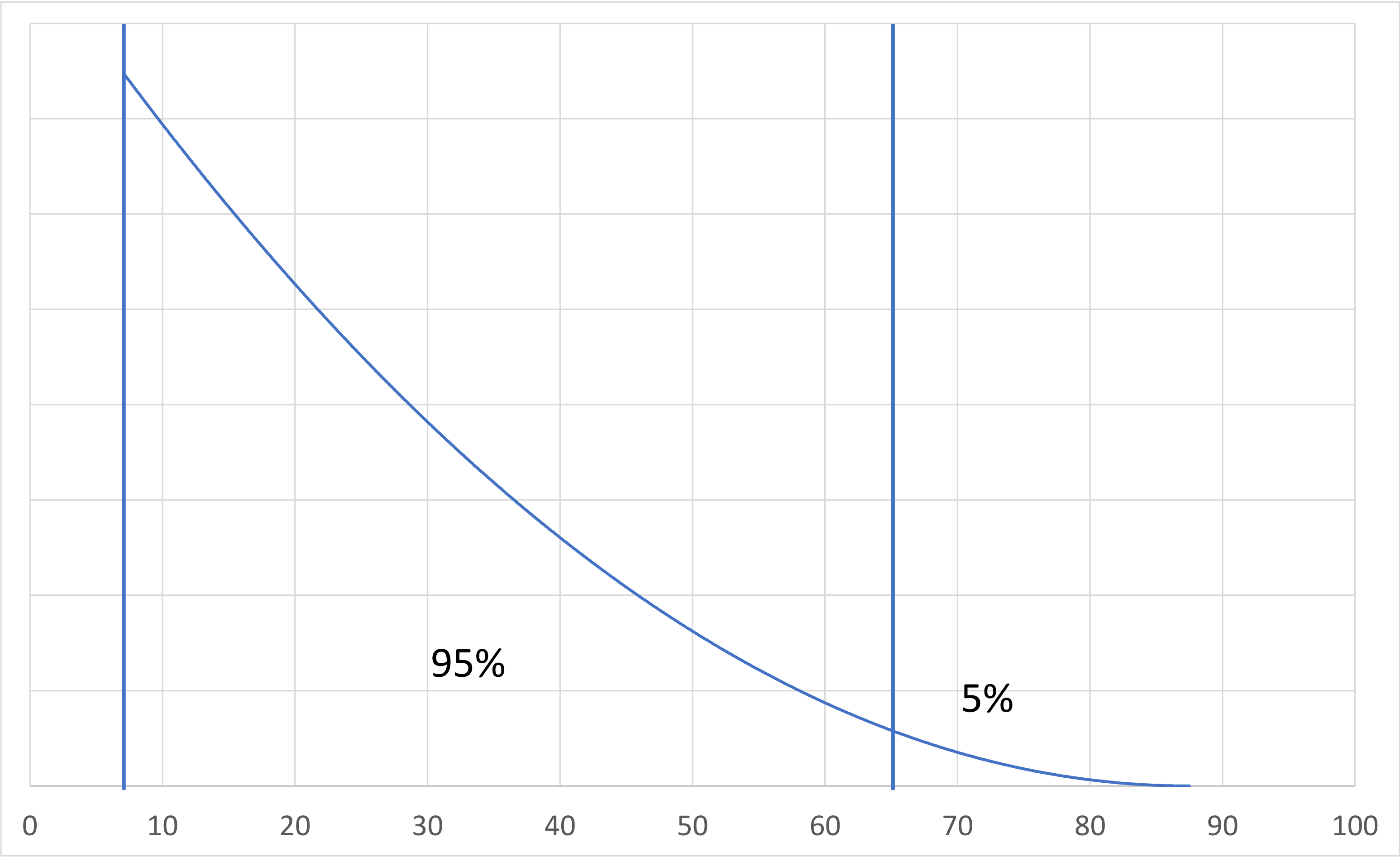
Най-тесен доверителен интервал

$$P\left[\pi = \frac{f + (k + 1)}{N}\right] - P\left(\pi = \frac{f + k}{N}\right)$$

$$= \frac{C_{N-n-(k+1)+m-2}^{m-2}}{C_{N-n+m-1}^{m-1}} - \frac{C_{N-n-k+m-2}^{m-2}}{C_{N-n+m-1}^{m-1}} = \cdots = -\frac{C_{N-n-k+m-3}^{m-3}}{C_{N-n+m-1}^{m-1}} < 0$$

Следователно $f(a) > f(b) \to (b - a)' > 0$

Най-тесен доверителен интервал



Доверителен интервал

$$F(x) = 1 - \left[\frac{1 - x - \frac{n}{N}(1-p)}{1 - \frac{n}{N}}\right]^{m-1} = P$$

$$x = p\frac{n}{N} + \left(1 - \frac{n}{N}\right)\left(1 - \sqrt[m-1]{1-P}\right)$$

$$P\left[p\frac{n}{N} \leq \pi \leq p\frac{n}{N} + \left(1 - \frac{n}{N}\right)\left(1 - \sqrt[m-1]{1-P}\right)\right] = P$$

Ширина на доверителния интервал

$$\left(1-\frac{n}{N}\right)\left(1-\sqrt[m-1]{1-P}\right)$$

Намаляваща спрямо $\frac{n}{N}$

Намаляваща спрямо m

Растяща спрямо P

Тест с емпирични данни

- Онлайн изследване, проведено от 16.04.2010 до 19.07.2010
- Брой посетители на сайта: 2876
- Брой уникални посетители на сайта: $N = 1985$
- Брой уникални посетители, попълнили анкетата: $n = 390$
- $\frac{n}{N} = 0,196$

Анкетна карта

- По каква система искате да избирате Вашите общински съветници?
- Пол
- Тип населено място
- Област
- Възрастова група
- Политическа ориентация

По каква система искате да избирате Вашите общински съветници?

Мажоритарна	$p = 0{,}362$	$P(0{,}071 \leq \pi \leq 0{,}579) = 0{,}95$
Пропорционална	$p = 0{,}056$	$P(0{,}011 \leq \pi \leq 0{,}519) = 0{,}95$
Пропорционална с преференция (с възможност да преподреждате партийната листа)	$p = 0{,}472$	$P(0{,}093 \leq \pi \leq 0{,}600) = 0{,}95$
Смесена (част пропорционално, част мажоритарно)	$p = 0{,}110$	$P(0{,}022 \leq \pi \leq 0{,}529) = 0{,}95$

Пол

Жена	$p = 0{,}333$	$P(0{,}065 \leq \pi \leq 0{,}829) = 0{,}95$
Мъж	$p = 0{,}667$	$P(0{,}131 \leq \pi \leq 0{,}894) = 0{,}95$

Тип населено място

Град	$p = 0{,}974$	$P(0{,}191 \leq \pi \leq 0{,}955) = 0{,}95$
Село	$p = 0{,}026$	$P(0{,}005 \leq \pi \leq 0{,}768) = 0{,}95$

Възрастова група

≤ 19	$p = 0{,}000$	$P(0{,}000 \leq \pi \leq 0{,}424) = 0{,}95$
20 – 29	$p = 0{,}213$	$P(0{,}042 \leq \pi \leq 0{,}465) = 0{,}95$
30 – 39	$p = 0{,}295$	$P(0{,}058 \leq \pi \leq 0{,}481) = 0{,}95$
40 – 49	$p = 0{,}285$	$P(0{,}056 \leq \pi \leq 0{,}479) = 0{,}95$
50 +	$p = 0{,}208$	$P(0{,}041 \leq \pi \leq 0{,}464) = 0{,}95$

Политическа ориентация

Дясно	$p = 0{,}637$	$P(0{,}110 \leq \pi \leq 0{,}743) = 0{,}95$
Ляво	$p = 0{,}104$	$P(0{,}018 \leq \pi \leq 0{,}651) = 0{,}95$
Център	$p = 0{,}260$	$P(0{,}045 \leq \pi \leq 0{,}678) = 0{,}95$

Заключение

- Ако непредставителните извадки са получени чрез възвратен подбор, те са абсолютно безполезни

- Ако непредставителните извадки са получени чрез безвъзвратен подбор, ползата от тях зависи от дела на извадката спрямо генералната съвкупност

    - Ако делът на извадката е пренебрежимо малък, тогава извадката е безполезна

    - Колкото е по-голям делът на извадката, толкова извадката е по-полезна

А сега накъде?

- Харалампиев, К. Нетрадиционен поглед върху традиционни статистически проблеми. Издателство „Балкани”, София, 2004
- Харалампиев, К. Анкетите в интернет страници – възможност за статистически изводи и интерпретиране на резултатите. Социологически проблеми, 3-4/2004
- Харалампиев, К. Телевизионните гласувания по телефона – проблемът за победителя. Социологически проблеми, 3-4/2005
- Харалампиев, К. Студентската оценка за преподаването – проблемът за точността на изводите. В: Социологията пред предизвикателството на различията. Юбилеен сборник, посветен на 30-годишнината на катедра „Социология”. Университетско издателство „Св. Климент Охридски“, София, 2009
- Харалампиев, К. Още една гледна точка към проблема за отказите при социологически изследвания. Социологически проблеми, 1-2/2012
- Харалампиев, К. Използване на бейсовска статистика за статистически изводи при непредварителни извадки (през примера на изследването на отпадащи студенти в бакалавърска степен на обучение във Философския факултет на Софийския университет “Св. Климент Охридски”). Реторика и комуникации, 36/2018