Laboratorium 6 – Estymacja

Niech $X_1, \dots, X_n$ zmienne losowe o takim samym rozkładzie normalnym $N(m, \sigma)$ reprezentującym pewną cechę elementów populacji.

1. Rozkład średniej arytmetycznej z próby

Średnia z próby $\bar{X} = \sum_{i=1}^{n} X_i$ jest zmienną losową o rozkładzie normalnym $N(m, \frac{\sigma}{\sqrt{n}})$, ponadto $D^2(\bar{X}) = \frac{\sigma^2}{n}$, $D(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

Standaryzowana zmienna $Z = \frac{\bar{X}-m}{\sigma}\sqrt{n}$ ma rozkład normalny $N(0,1)$.

Statystyka $t = \frac{\bar{X}-m}{S}\sqrt{n-1}$ gdzie $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ ma rozkład t-Studenta o n – 1 stopni swobody.

2. Rozkład wariancji z próby

Mamy dwie statystyki: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, $\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ oraz $S^2 = \hat{S}^2\,\frac{n}{n-1}$. Dla tych statystyk mamy: $E(S^2) = \frac{n-1}{n}\,\sigma^2$, $D^2(S^2) = \frac{2(n-1)}{n^2}\sigma^4$; $E(\hat{S}^2) = \sigma^2$, $D^2(\hat{S}^2) = \frac{2}{n-1}\sigma^4$.

3. Rozkład różnicy średnich. Dane są dwie populacje generalne o niezależnych rozkładach normalnych $N(m_1, \sigma_1), N(m_2, \sigma_2)$. Z populacji tych wylosowano dwie próby o licznościach n_1, n_2. Statystyka różnicy średnich tych prób $Z = \bar{X}_1 - \bar{X}_2$ ma również rozkład normalny

$$N\left(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Zmienna standaryzowana dla Z ma postać:

$$U = \frac{\bar{X}_1 - \bar{X}_2 - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Jeżeli $\sigma_1 = \sigma_2$ choć nie są znane, to studentyzowana zmienna

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (m_1 - m_2)}{\sqrt{\frac{n_1(S_1)^2 + n_2(S_2)^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

ma rozkład t-Studenta o $n_1 + n_2 - 2$ stopni swobody.

4. Rozkład ilorazu wariancji. Dane są dwie populacje generalne o niezależnych rozkładach normalnych $N(m_1, \sigma_1), N(m_2, \sigma_2)$ z jednakowymi odchyleniami standardowymi $\sigma_1 = \sigma_2 = \sigma$. Z populacji tych wylosowano dwie próby o licznościach n_1, n_2. Dla każdej z nich obliczono estymator wariancji

$$\left(\hat{S}_i\right)^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left((X_i)_j - \bar{X}_i\right)^2, \quad i = 1,2$$

Statystyka

$$F = \frac{\left(\hat{S}_1\right)^2}{\left(\hat{S}_2\right)^2}$$

ma rozkład F-Snedecora o $n_1 - 1$ stopniach swobody.

5. Rozkład frakcji. Ma zastosowanie przy badaniu cechy jakościowej, niemierzalnej (numerycznie). Populację dzielimy na klasę posiadającą te cechę i drugą nie posiadającą. Rozkład cechy w populacji jest dwupunktowy

$$\Pr(X = x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases}$$

Niech $X_1, \dots, X_n$ zmienne losowe o takim rozkładzie (próba prosta). Statystyka

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

gdzie *m* jest liczbą elementów o wartości 1 w *n*-elementowej próbie. Dla parametru rozkładu struktury p stosujemy estymator

$$\hat{p} = \frac{m}{n}$$

wtedy $E(\hat{p}) = p,\ \ D^2(\hat{p}) = \frac{p(1-p)}{n}$. Dla dużych *n* > 100 rozkład wielomianowy $\hat{p}$ może być zastąpiony rozkładem $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

6. Rozkład różnicy frakcji. Badamy rozkład cechy niemierzalnej o rozkładzie dwupunktowym. Pobrano dwie duże próby losowe o licznościach $n_1, n_2 > 100$. Różnica estymatorów $\hat{p}_i = \frac{m_i}{n_i}$, $i = 1,2$ ma rozkład asymptotycznie normalny $N\left(p_1 - p_2,\ \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$.

Zadanie 1. Miesięczne wydatki studenta na artykuły piśmienne mają rozkład N(33,8). Ustal prawdopodobieństwo, że w losowo pobranej próbie o liczności *n* = 16 średnia arytmetyczna nie przekroczy 35.

Rozwiązanie:
Średnia w próbie *n* = 16 ma rozkład $N(33, \frac{8}{\sqrt{16}})$. Stosując procedurę biblioteczną R trzeba policzyć Pr $(X \leq 35)$ z tego rozkładu.

```
stats::pnorm(35,mean=33,sd=8/sqrt(16),lower.tail=TRUE)
```

Niech $X_1, \dots, X_n$ próba losowa z populacji, Q parametr rozkładu cechy X w populacji, Estymator Q parametru X, to statystyka $Z_n = f(X_1, \dots, X_n)$.
Estymator Z_n dla Q nazywamy **nieobciążonym**, jeżeli

$$E(Z_n) = Q.$$

Wielkość obciążenia, to

$$b(Z_n) = E(Z_n) - Q.$$

Estymator Z_n dla Q nazywamy **zgodnym**, jeżeli

$$\lim_{n\to+\infty} \Pr(|Z_n - Q| < \varepsilon) = 1 \quad \forall \varepsilon > 0.$$

Estymator Z_n dla Q nazywamy **najefektywniejszym**, jeżeli

$$\forall Z_n^* \;\; e(Z_n) = \frac{D^2(Z_n^*)}{D^2(Z_n)} \geq 1$$

gdzie Z_n^* jest innym estymatorem dla Q.
Estymator Z_n dla Q nazywamy **wystarczającym**, jeżeli wykorzystuje wszystkie dostępne w próbie $X_1, \dots, X_n$ informacje o parametrze Q.

Parametr	**Estymator**	**Własności**
$E(X) = m$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	nieobciążony, zgodny, efektywny
Prawdopodobieństwo (wskaźnik struktury, frakcja cechy wyróżnionej) $p(X)$	$\bar{p} = \frac{m}{n}$	nieobciążony, zgodny, efektywny
$D^2(X)$	$S_1^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - m)^2$	nieobciążony, zgodny, efektywny (wymaga znajomości m)
	$S_2^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$	zgodny
	$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n}{n-1} S_2^2$	nieobciążony, zgodny
$D(x)$	$S_2 = \sqrt{S_2^2}$	zgodny
	$\hat{S} = \sqrt{\hat{S}^2}$	zgodny
$E(X^k)$	$A_k = \frac{1}{n}\sum_{i=1}^{n} (X_i)^k$	nieobciążony, zgodny
$E((X-m)^k)$	$B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k$	zgodny

Estymacja parametryczna, to szacowanie nieznanych parametrów populacji o ustalonym typie rozkładu.

Estymacja nieparametryczna, to szacowanie nieznanej postaci funkcyjnej rozkładu populacji (inaczej aproksymacja rozkładu).

Estymacja punktowa, to podanie jednej wartości parametru rozkładu cechy (być może wartości wektorowej, tupli).

Estymacja przedziałowe (dla cechy skalarnej Q) polega na znalezieniu jednego z przedziałów $[a, b] \subset \mathbb{R}$ spełniającego warunek

$$\Pr(a < Q < b) = 1 - \alpha$$

gdzie $1 - \alpha \in (0,1)$ jest pewnym, zwykle bardzo dużym prawdopodobieństwem zwanym poziomem ufności.
Uwaga. Bardzo duże poziomy ufności powodują wzrost długości przedziału ufności, co naprawdę zmniejsza przydatność oszacowania.

7. Estymacja wartości oczekiwanej

7.1. Estymacja punktowa

Metoda momentów. Pobieramy próbę losową prostą $X_1, \dots, X_n$. Stosujemy estymator (najlepiej $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$). Za najbardziej prawdopodobny błąd uznajemy estymator wariancji $D(\bar{X})$. Zatem

$$E(X) \cong \bar{x} \pm \frac{s}{\sqrt{n}}\,, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Można również zastosować estymatory oparte na **funkcji największej wiarygodności**, posługując się funkcjami z pakietu R.

Funkcja `fitdistr()` z pakietu MASS

```
fitdistr(x,densfun,start, …)
```

gdzie – `x` wektor danych do estymacji, `densfun` – string określający typ gęstości rozkładu (`'beta','cauchy','chi-squared','exponential','gamma','geometric','lognormal','logistic', 'negative binomial', 'normal', 'Poisson','t','wibull'`), stats – parametry określające rozkład, który ma być estymowany, jeżeli nie jest to któryś ze standardowych.

Można również używać funkcji `fitdist()` i `descdist()` z pakietu `fitdistrplus`.

7.2. Estymacja przedziałowa

Model I. Niech cecha X ma w populacji rozkład $N(m, \sigma)$ przy czym m jest nieznane a σ znane.

$$\Pr\left(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} < m < \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

gdzie Z ma rozkład normalny $N(0,1)$, oraz $\Pr(-z_\alpha < Z < z_\alpha) = 1 - \alpha$, natomiast $1 - \alpha$ jest poziomem ufności. Precyzję oszacowani określa się wzorem

$$B(\bar{x}) = \frac{z_\alpha\, \sigma}{\bar{x}\sqrt{n}}\, 100\%$$

Wartość z_α jest kwantylem rzędu $q = 1 - \frac{\alpha}{2}$ rozkładu $N(0,1)$, zatem można ją obliczyć funkcją

```
zalpha = stats::qnorm(q, mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)
```
(por. Górecki; Podsatwy statystyki z przykładami w R, BTC 2011)

Zadanie 2. Załóżmy, że zawartość tłuszczu w mleku jest zmienną losową o rozkładzie $N(m, 0.2)$. Wykonano n = 25 niezależnych pomiarów i otrzymano średnią $\bar{x} = 3.15\%$. Zbuduj przedział ufności dla poziomu ufności 0.95.

Model II. Niech cecha X ma w populacji rozkład $N(m, \sigma)$ przy czym m, σ są nieznane.

$$\Pr\left(\bar{x} - t_\alpha \frac{s}{\sqrt{n-1}} < m < \bar{x} + t_\alpha \frac{s}{\sqrt{n-1}}\right) = 1 - \alpha, \qquad s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

lub

$$\Pr\left(\bar{x} - t_\alpha \frac{\hat{s}}{\sqrt{n-1}} < m < \bar{x} + t_\alpha \frac{\hat{s}}{\sqrt{n-1}}\right) = 1 - \alpha, \qquad \hat{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

gdzie t_α jest wartością zmiennej o rozkładzie t-Studenta o n - 1 stopniach swobody spełniającą

$$\Pr(-t_\alpha < t < t_\alpha) = 1 - \alpha$$

Precyzja oszacowania jest równa

$$B(\bar{x}) = \frac{t_\alpha \hat{s}}{\bar{x}\sqrt{n}}\ 100\%$$

Wartość t_α jest kwantylem rzędu $q = 1 - \frac{\alpha}{2}$ rozkładu t(n-1), zatem można ją obliczyć funkcją
```
talpha = stats::qt(q, df=n-1, lower.tail = TRUE, log.p = FALSE)
```
(por. Górecki; Podsatwy statystyki z przykładami w R, BTC 2011)

Zadanie 3. Chcemy przy poziomie ufności 0.95 oszacować średni wiek lekarzy pracujących w wiejskich ośrodkach zdrowia. Wylosowano próbę n = 25 osób dla których otrzymano $\bar{x} = 45$ lat przy s = 11 lat.

Model III. Niech cecha X ma w populacji rozkład $N(m, \sigma)$ lub inny rozkład o skończonej średniej i wariancji, przy czym m, σ są nieznane, wtedy

$$\Pr\left(\bar{x} - z_\alpha \frac{\hat{s}}{\sqrt{n}} < m < \bar{x} + z_\alpha \frac{\hat{s}}{\sqrt{n}}\right) = 1 - \alpha, \qquad \hat{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

przy czym ponownie Z ma rozkład normalny $N(0,1)$, oraz $\Pr(-z_\alpha < Z < z_\alpha) = 1 - \alpha$, natomiast $1 - \alpha$ jest poziomem ufności. Ponadto precyzja oszacowania dana jest wzorem

$$B(\bar{x}) = \frac{z_\alpha \hat{s}}{\bar{x}\sqrt{n}}\ 100\%$$

Zadanie 4. Chcemy oszacować średni staż pracy pracowników X w drukarni na poziomie ufności 0.96. W tym celu ustalono próbę n = 100 pracowników, dla których $\bar{x} = 5.4$ natomiast $\hat{s} = 1.7$ lat.

8. Estymacja przedziałowa wariancji i odchylenia standardowego

Dla wariancji stosujemy dwa estymatory:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad \hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Model I. Niech cecha X ma w populacji rozkład $N(m, \sigma)$ przy czym m, σ są nieznane. Losujemy małą $n \leq 30$ próbę losową. Wtedy

$$\Pr\left(\frac{ns^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{ns^2}{\chi^2_{\frac{\alpha}{2},n-1}}\right) = 1-\alpha, \qquad s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

lub

$$\Pr\left(\frac{(n-1)\hat{s}^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)\hat{s}^2}{\chi^2_{\frac{\alpha}{2},n-1}}\right) = 1-\alpha, \qquad \hat{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

gdzie $\chi^2_{\frac{\alpha}{2},n-1}, \chi^2_{1-\frac{\alpha}{2},n-1}$ są wartościami zmiennej losowej o rozkładzie χ^2 dla $n-1$ stopni swobody. Nie jest wyjaśnione, jak wartości zmiennej o tym rozkładzie zależą od parametru α?

Wartości $\chi^2_{\frac{\alpha}{2},n-1}, \chi^2_{1-\frac{\alpha}{2},n-1}$ są kwantylami rzędu $q1 = \frac{\alpha}{2}$, $q2 = 1 - \frac{\alpha}{2}$ rozkładu chi-kwadrat o (n-1) stopniach swobody, zatem można ją obliczyć funkcją

```
chji1 = stats::qchisq(q1, df=n-1, lower.tail = TRUE, log.p =
FALSE)
chji2 = stats::qchisq(q2, df=n-1, lower.tail = TRUE, log.p =
FALSE)
```

(Być może w drugim wywołaniu funkcji powinno być `lower.tail = FALSE` ???)

(por. Górecki; Podsatwy statystyki z przykładami w R, BTC 2011)

Uwaga. Jeżeli $\Pr(a < \sigma^2 < b) = 1-\alpha$ jest przedziałem ufności dla wariancji, to $\Pr\left(\sqrt{a} < \sigma < \sqrt{b}\right) = 1-\alpha$ jest przedziałem ufności dla odchylenia standardowego.

Uwaga. Ponieważ przedziały ufności dla σ^2 i σ nie są symetryczne względem wartości estymatora nie istnieje dla nich względny miernik precyzji.

Zadanie 5. Mamy ocenić przy poziomie ufności 0.90 zróżnicowanie średnicy pni drzew w lesie na podstawie n = 25 elementowej próby dla której otrzymano $\bar{x} = 37.3$ cm, $s^2 = 13.5$ cm^2.

Zakładamy, że średnica pni drzew ma rozkład normalny $N(m,\sigma)$ oraz, że miernikiem zróżnicowania jest wariancja σ^2.

Model II. Niech cecha X ma w populacji rozkład $N(m,\sigma)$ przy czym m, σ są nieznane. Losujemy dużą $n > 30$ próbę losową. Wtedy możemy korzystać z aproksymacji rozkładu χ^2 rozkładem normalnym $N\left(n-1,\sqrt{2(n-1)}\right)$. Wtedy

$$\Pr\left(\frac{s}{1+\frac{z_\alpha}{\sqrt{2n}}} < \sigma < \frac{s}{1-\frac{z_\alpha}{\sqrt{2n}}}\right) = 1-\alpha, \qquad s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

przy czym ponownie Z ma rozkład normalny $N(0,1)$, oraz $\Pr(-z_\alpha < Z < z_\alpha) = 1-\alpha$, natomiast $1-\alpha$ jest poziomem ufności. Ponadto precyzja oszacowania dana jest wzorem

$$B(s) = \frac{|z_\alpha|}{\sqrt{2n}}\ 100\%$$

Zadanie 6. W próbie 450 samochodów jednego typu przeprowadzono badanie zużycia paliwa na 100 km. Okazało się, że odchylenie standardowe w tej próbie było równe *s* = 0.8 litra/100 km. Wyznaczyć przedział ufności dla tej zmiennej dla poziomu ufności 0.99.

9. Estymacja przedziałowa wskaźnika struktury dla dużej próby (n > 100)

Stosujemy estymator

$$\bar{p} = \frac{m}{n}$$

gdzie *m* – ilość elementów posiadających wyróżnioną cechę w próbie, *n* – liczność próby.

$$\Pr\left(\frac{m}{n} - z_\alpha\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}} < p < \frac{m}{n} + z_\alpha\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}\right) = 1-\alpha$$

przy czym ponownie Z ma rozkład normalny $N(0,1)$, oraz $\Pr(-z_\alpha < Z < z_\alpha) = 1-\alpha$, natomiast $1-\alpha$ jest poziomem ufności. Ponadto precyzja oszacowania dana jest wzorem

$$B\left(\frac{m}{n}\right) = \frac{z_\alpha}{\frac{m}{n}}\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}\ 100\%$$

Zadanie 7. Z pośród 10 000 pracowników korporacji wylosowano próbę prostą o liczności *n* = 200. Pytano ich o plany zmiany pracy. Tylko 20 pracowników zamierza opuścić firmę. Dla poziomu ufności 0.90 wyznaczyć przedział ufności dla wskaźnika struktury pracowników, którzy chcą opuścić firmę.

10. minimalna liczebność próby

Uwaga. Zwiększanie poziomu ufności przy stałym *n* powoduje wydłużenie przedziału ufności, w konsekwencji gorsze oszacowanie. Zwiększenie *n* przy stałym poziomie ufności powoduje skrócenie długości przedziału ufności. Próba odpowiedzi na pytanie odwrotne, tj. jaka ma być liczność próby, aby przy zadanym poziomie ufności uzyskać założoną długość przedziału ufności nie zawsze jest możliwa.

Dla szacowania wielkości próby możemy wykorzystać funkcję `nsize()` z pakietu `PASWR`.

`nsize(b, sigma = NULL, p = 0.5, conf.level = 0.95, type = "mu")`

- `b` - pożądane ograniczenie (chyba długości przedziału?), `sigma` – odchylenie standardowe populacji, niekonieczne jeżeli mamy typ „chi", `p` – estyma ta dla proporcji sukcesu populacji, niekonieczna gdy typ „mu", `conf.level` – poziom ufności $(1-\alpha)$, `type` – string „mu" albo „pi" wskazujący na estymowany parametr.

10.1. Minimalna wielkość próby przy szacowaniu średniej

Model I. Cecha w populacji ma rozkład normalny przy znanym odchyleniu standardowym σ, to długość przedziału ufności jest równa $2d = 2z_\alpha \frac{\sigma}{\sqrt{n}}$. Wielkość d możemy traktować jako maksymalny błąd szacunku cechy. Chcąc utrzymać założone d i wynikającą z przyjętego poziomu ufności wartość zmiennej z_α musimy spełnić

$$n \geq \left\lceil \frac{z_\alpha^2 \sigma^2}{d^2} \right\rceil$$

Zadanie 8. Mamy ustalić minimalną liczebność próby dla oszacowania średniej wzrostu noworodków o rozkładzie $N(m, 1.5)$. Zakładamy maksymalny błąd szacunku *d* = 0.5 cm oraz poziom ufności 0.99.

Model II. Cecha w populacji ma rozkład normalny o nieznanych parametrach. Długość przedziału ufności ustalamy korzystając ze wzoru

$$n - \frac{t_\alpha^2 \hat{s}^2}{d^2}$$

gdzie $\hat{s} = \sqrt{\frac{1}{n_0 - 1}\sum_{i=1}^{n_0}(x_i - \bar{x})^2}$ jest statystyką obliczoną z małej próby pilotażowej o liczności n_0, t_α jest wartością zmiennej losowej o rozkładzie *t*-Studenta o $n_0 - 1$ stopniach swobody spełniającą $\Pr(-t_\alpha < t < t_\alpha) = 1 - \alpha$.

Wybieramy próbę pilotażową o liczności n_0. Wyznaczamy dla niej $\bar{x}$ i $\hat{s}$. Dla α i $n_0 - 1$ wyznaczamy t_α oraz obliczamy $n = \frac{t_\alpha^2 \hat{s}^2}{d^2}$. Jeżeli $n \leq n_0$ to uznajemy próbę pilotażową za satysfakcjonującą. Jeżeli $n > n_0$ do dolosowujemy $n - n_0$ elementów próby.

Zadanie 9. Chcemy oszacować średni wzrost uczniów w klasach 5-tych przy maksymalnym błędzie szacunku *d* = 5 cm i poziomie ufności 0.95. Wybieramy próbę pilotażową $n_0 = 10$ uczniów. Uzyskujemy dla niej $\bar{x} = 142$ cm i $\hat{s} = 169$ cm^2.

10.1. Minimalna wielkość próby przy szacowaniu wskaźnika struktury

Długość przedziału ufności jest równa w tym przypadku

$$2z_{\alpha}\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}$$

Chcąc ograniczyć ją przez 2*d* otrzymujemy

$$n \geq \left\lceil \frac{z_{\alpha}^2 \frac{m}{n}\left(1-\frac{m}{n}\right)}{d^2} \right\rceil$$

W powyższym wzorze $\frac{m}{n}$ możemy traktować jako obserwowany estymator wskaźnika struktury $\bar{p}$ uzyskany w badaniu pilotażowym.

Zadanie 10. Należy ustalić minimalną wielkość próbki dla określenia odsetki zgonów w mieście przy poziomie ufności 0.95 i maksymalny błędzie szacunku *d* = 1%. W wyniku badań pilotażowych otrzymano odsetek zgonów równy 0.45%.

W sytuacji, w której nie mamy badań pilotażowego dla estymacji $\bar{p}$ możemy zamiast niej przyjąć najniekorzystniejszy wariant $\bar{p} = 0.5$ dla której $\frac{m}{n}\left(1-\frac{m}{n}\right)$ przyjmuje maksymalną wartość 0.25, wtedy

$$n \geq \left\lceil \frac{0.25\, z_{\alpha}^2}{d^2} \right\rceil$$

Zadanie 11. Chcemy zbadać odsetek rodzin w dużym mieście, które chcą korzystać z Internetu nie mając pilotażowej wartości estymatora tej cechy. Na ilu rodzinach należy przeprowadzić ankietę, aby przy poziomie ufności 0.90 otrzymać przedział ufności odpowiadający 8% dokładności?