2 Statystyka: estymacja

Model statystyczny

- próba: $X = (X_1, \dots, X_n)$ – ciąg zmiennych losowych – wynik eksperymentu, pomiaru, obserwacji,
- $\mathcal{X}$ – przestrzeń próby – zbiór wszystkich możliwych wartości X,
- $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ – rodzina rozkładów prawdopodobieństwa na przestrzeni prób $\mathcal{X}$,
- θ – parametr, Θ – zbiór możliwych wartości parametru θ,
- Próba prosta (z rozkładu $\mathbb{P}_\theta$): $(X_1, X_2, \dots, X_n)$ – niezależne zmienne losowe o tym samym rozkładzie ($\mathbb{P}_\theta$).
- $(\mathcal{X}, \mathcal{P})$ – model statystyczny (przestrzeń statystyczna),
- $f\colon \mathcal{X} \to \mathbb{R}$ – statystyka z próby – zmienna losowa będąca funkcją obserwowanych w próbie zmiennych losowych (nie zależy bezpośrednio od θ).

Ważne rozróżnienie:

- $X_1, X_2, \dots, X_n$ – próba (zmienne losowe),
- statystyki z próby (zmienne losowe):
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$
- $x_1, x_2, \dots, x_n$ – realizacja próby (wartości przyjęte przez zmienne losowe – liczby),
- oceny statystyk (liczby):
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

2.1 Estymacja

- estymacja parametryczna – szacowanie nieznanych wartości parametrów rozkładu cechy statystycznej w populacji generalnej,
- estymacja nieparametryczna – szacowanie nieznanego rozkładu badanych cech w populacji generalnej,
- estymacja punktowa – za ocenę wartości przyjmujemy jedną wartość (dodając błąd szacunku),
- estymacja przedziałowa – wyznaczamy przedział, w którym z dużym prawdopodobieństwem znajduje się wartość szacowanego parametru.

Estymator

- Estymator to statystyka, która służy oszacowaniu parametru(ów) rozkładu.
- Estymatorem parametru θ rozkładu zmiennej losowej X nazywamy statystykę $\hat{\theta}_n = f_n(X_1, \dots, X_n)$, której rozkład prawdopodobieństwa zależy od θ.
- Liczbę $f_n(x_1, \dots, x_n)$ jaką przyjmuje estymator $\hat{\theta}_n$ dla realizacji próby $(x_1, \dots, x_n)$ nazywamy oceną parametru θ.
- Indeks dolny n w zapisie $\hat{\theta}_n$ oznacza, że estymator wyznaczamy w oparciu o próbę n-elementową.

Pożądane cechy estymatorów

- Liczbę $B(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n - \theta)$ nazywamy obciążeniem estymatora,
- Estymator nazywamy nieobciążonym, jeśli
$$B(\hat{\theta}_n) = 0, \quad \text{czyli} \quad \mathbb{E}(\hat{\theta}_n) = \theta.$$
- Estymator nazywamy asymptotycznie nieobciążonym, jeśli
$$\lim_{n\to\infty} B(\hat{\theta}_n) = 0, \quad \text{czyli} \quad \lim_{n\to\infty} \mathbb{E}(\hat{\theta}_n) = \theta.$$
- Estymator nazywamy zgodnym, jeśli (*zbieżność według prawdopodobieństwa – stochastyczna*)
$$\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta| < \varepsilon) = 1 \quad \text{dla każdego } \varepsilon > 0.$$
- Jeśli estymator jest zgodny, to jest asymptotycznie nieobciążony.
- Jeśli estymator jest asymptotycznie nieobciążony i jego wariancja maleje wraz ze wzrostem liczebności próby do zera, to jest zgodny.

Twierdzenie. *Prawo wielkich liczb Bernoulliego Jeśli k oznacza liczbę sukcesów w n próbach Bernoulliego, to*
$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{k}{n} - p\right| < \varepsilon\right) = 1 \quad \textit{dla każdego } \varepsilon > 0,$$
gdzie p jest prawdopodobieństwem sukcesu w pojedynczym doświadczeniu.

Twierdzenie. *Prawo wielkich liczb Chinczyna Jeśli (X_n) jest ciągiem niezależnych zmiennych losowych o tym samym rozkładzie i skończonej wartości oczekiwanej $\mathbb{E}(X_1) = \mu$, to*
$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| < \varepsilon\right) = 1 \quad \textit{dla każdego } \varepsilon > 0.$$

Przykład (Średnia i wariancja z próby)**.** Niech $(X_1, X_2, \ldots, X_n)$ będzie próbą prostą z (dowolnego) rozkładu o parametrach: $\mathbb{E}(X_1) = \mu$, $\mathbb{D}^2(X_1) = \sigma^2$. Jako estymator parametru μ rozważmy średnią arytmetyczną z próby:
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$
Z niezależności oraz własności operatorów $\mathbb{E}$ i $\mathbb{D}^2$ mamy
$$\begin{aligned}
\mathbb{E}(\bar{X}) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n} n\mu = \mu, \\
\mathbb{D}^2(\bar{X}) &= \mathbb{D}^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{D}^2(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}, \\
\mathbb{D}(\bar{X}) &= \frac{\sigma}{\sqrt{n}}.
\end{aligned}$$
Dalej, rozważmy estymator parametru σ^2 dany wzorem:
$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Okazuje się, że jest to estymator obciążony (choć asymptotycznie nieobciążony):

$$\begin{aligned}
\mathbb{E}(S^2) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right) = \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i^2\right) - \frac{1}{n^2}\mathbb{E}\left(\sum_{i,j=1}^{n} X_i X_j\right) = \\
&= \frac{1}{n}n(\sigma^2+\mu^2) - \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}\left(X_i^2\right) + \sum_{i,j=1,i\neq j}^{n}\mathbb{E}\left(X_i X_j\right)\right) = \\
&= \sigma^2+\mu^2 - \frac{1}{n^2}\left(n(\sigma^2+\mu^2)+(n^2-n)\mu^2\right) = \\
&= \sigma^2+\mu^2 - \frac{1}{n^2}\left(n\sigma^2+n^2\mu^2\right) = \left(1-\frac{1}{n}\right)\sigma^2 = \frac{n-1}{n}\sigma^2.
\end{aligned}$$

Jednakże dzięki prostej „poprawce” uzyskujemy estymator nieobciążony:

$$\begin{aligned}
\hat{S}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{n}{n-1}S^2, \\
\mathbb{E}(\hat{S}^2) &= \frac{n}{n-1}\mathbb{E}(S^2) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2.
\end{aligned}$$

Uwaga: jeśli parametr μ jest znany, dla statystyki $S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ mamy jednakże

$$\begin{aligned}
\mathbb{E}(S_\mu^2) &= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i^2 - 2\mu\sum_{i=1}^{n} X_i + n\mu^2\right) = \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - 2\mu\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) + \mu^2 = \\
&= \frac{1}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2) - 2\mu^2 + \mu^2 = \sigma^2+\mu^2-\mu^2 = \sigma^2.
\end{aligned}$$

Pożądane cechy estymatorów c.d.

- Wariancja estymatora:
$$\mathbb{D}^2(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))^2.$$
- Błąd średniokwadratowy estymatora:
$$MSE(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n - \theta)^2$$
- Mamy
$$MSE(\hat{\theta}_n) = \mathbb{D}^2(\hat{\theta}_n) + [B(\hat{\theta}_n)]^2$$
- Jeśli estymator jest nieobciążony, to $MSE(\hat{\theta}_n) = \mathbb{D}^2(\hat{\theta}_n)$.
- $\mathbb{D}(\hat{\theta}_n)$ nazywamy wówczas średnim (standardowym) błędem szacunku parametru θ, a $\mathbb{D}(\hat{\theta}_n)/\theta$ jest względnym błędem szacunku.
- Estymator nazywamy najefektywniejszym w danej klasie estymatorów, jeśli ma w tej klasie najmniejszą wariancję. Zwykle efektywność rozważamy w klasie estymatorów nieobciążonych.

- Estymator nazywamy efektywnym w sensie Rao-Cramera jeśli jest nieobciążony i realizuje dolne ograniczenie w nierówności Rao-Cramera:

$$\mathbb{D}^2(\hat{\theta}_n) \geqslant \left(n\mathbb{E}\left[\left(\frac{\partial \ln f(x;\theta)}{\partial \theta} \right)^2 \right] \right)^{-1},$$

  gdzie $f(x;\theta)$ jest funkcją gęstości lub funkcją prawdopodobieństwa populacji generalnej.

Przykład. Obliczmy prawą stronę nierówności Rao-Cramera dla rozkładu normalnego:

$$\begin{aligned}
X_i &\sim N(\mu, \sigma^2),\\
f(x;\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right),\\
\ln f(x;\mu) &= -\ln(\sqrt{2\pi\sigma^2}) - \frac{(x-\mu)^2}{2\sigma^2},\\
\frac{\partial \ln f(x;\mu)}{\partial \mu} &= \frac{2(x-\mu)}{2\sigma^2} = \frac{x-\mu}{\sigma^2},\\
\mathbb{E}\left[\left(\frac{\partial \ln f(x;\mu)}{\partial \mu} \right)^2 \right] &= \mathbb{E}\left[\left(\frac{x-\mu}{\sigma^2} \right)^2 \right] = \frac{\mathbb{E}((x-\mu)^2)}{\sigma^4} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2},\\
\left(n\mathbb{E}\left[\left(\frac{\partial \ln f(x;\mu)}{\partial \mu} \right)^2 \right] \right)^{-1} &= \frac{\sigma^2}{n}.
\end{aligned}$$

Stąd (i z obliczeń w poprzednim przykładzie) wynika, że średnia arytmetyczna z próby prostej o rozkładzie normalnym, jako estymator parametru μ, jest estymatorem efektywnym w sensie Rao-Cramera.

Klasyczne metody uzyskiwania estymatorów

- metoda najmniejszych kwadratów,
- metoda momentów,
- metoda największej wiarygodności,
  - Funkcją wiarygodności próby nazywamy wyrażenie:

    $$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i;\theta).$$

    (wartości z próby są ustalone, a parametr θ jest zmienną).
  - Za $\hat{\theta}$ przyjmujemy wielkość maksymalizującą funkcję wiarygodności (lub jej logarytm),
  - Przy dość ogólnych założeniach estymatory MNW są zgodne, asymptotycznie normalne, asymptotycznie nieobciążone i asymptotycznie najefektywniejsze.

Własności rozkładu normalnego Jeśli $X_1, X_2, \dots, X_n$ są niezależne o rozkładach normalnych: $X_i \sim N(\mu_i, \sigma_i^2)$, to:

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2 \right).$$

Jeśli $X_1, X_2, \ldots, X_n$ jest próbą prostą z rozkładu normalnego $N(\mu, \sigma^2)$, to:

$$\sum_{i=1}^{n} X_i \sim N\left(n\mu, n\sigma^2\right),$$
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$
$$\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1).$$

Rozkłady t-Studenta, χ^2 oraz F Jeśli $X_1, X_2, \ldots, X_n$ jest próbą prostą z rozkładu normalnego $N(\mu, \sigma^2)$, to:

$$\frac{nS^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1}, \qquad \text{(rozkład chi kwadrat)}$$
$$\frac{\bar{X}-\mu}{S}\sqrt{n-1} = \frac{\bar{X}-\mu}{\hat{S}}\sqrt{n} \sim t_{n-1}, \qquad \text{(rozkład t-Studenta)}$$

Jeśli $X_1, \ldots, X_{n_1}$ oraz $Y_1, \ldots, Y_{n_2}$ są niezależnymi próbami prostymi z rozkładu normalnego, odpowiednio: $N(\mu_1, \sigma^2)$ i $N(\mu_2, \sigma^2)$ (σ^2 jest nieznane, ale takie samo w obu rozkładach!), to (rozkład F, F-Sndecora lub Fishera-Snedecora):

$$\frac{\hat{S}_X^2}{\hat{S}_Y^2} \sim F_{n_1-1, n_2-1}.$$

Przypomnienie: Centralne Twierdzenie Graniczne Lindeberga-Levy'ego Jeśli $(X_n)_{n\in\mathbb{N}}$ jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie, $\mathbb{E}(X_1) = \mu$, $\mathbb{D}^2(X_1) = \sigma^2 < \infty$, to

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n\sigma^2}} \leqslant x\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\, dt.$$

Inaczej mówiąc:

$$Y_n = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \quad \xrightarrow[n\to\infty]{D} \quad N(0,1),$$

– ciąg dystrybuant (F_{Y_n}) zmiennych losowych (Y_n) zbiega do dystrybuanty rozkładu normalnego standardowego.

2.2 Estymacja przedziałowa

Jerzy Spława-Neyman (1894-04-16 – 1981-08-05)

- Cecha X ma w populacji rozkład z nieznanym parametrem θ.
- Na podstawie wylosowanej z tej populacji próby $(X_1, \ldots, X_n)$ wyznaczamy

  $$\underline{\theta} = \underline{\theta}(X_1, \ldots, X_n), \qquad \overline{\theta} = \overline{\theta}(X_1, \ldots, X_n),$$

  aby dla przyjętego prawdopodobieństwa $1-\alpha$ zachodził warunek

  $$\mathbb{P}\left(\underline{\theta}(X_1, \ldots, X_n) < \theta < \overline{\theta}(X_1, \ldots, X_n)\right) = 1-\alpha.$$

- Losowy przedział $(\underline{\theta}, \overline{\theta})$ nazywamy przedziałem ufności parametru θ.
- Liczbę $1-\alpha$ nazywamy współczynnikiem (poziomem) ufności.
- Długość przedziału ufności $\overline{\theta} - \underline{\theta}$ określa dokładność estymacji. Zależy nam na największej dokładności – szukamy najkrótszych przedziałów ufności.

Przedział ufności dla dowolnego rozkładu Uniwersalny (dla dowolnego rozkładu) przedział ufności dla wartości oczekiwanej otrzymujemy z nierówności Czebyszewa:

$$\mathbb{P}(|X - \mathbb{E}(X)| < \varepsilon) \geqslant 1 - \frac{\mathbb{D}^2(X)}{\varepsilon^2}.$$

Stąd dla $1 - \alpha = 1 - \frac{\mathbb{D}^2(X)}{\varepsilon^2} \iff \varepsilon = \frac{\mathbb{D}(X)}{\sqrt{\alpha}}$ mamy

$$\mathbb{P}\left(X - \frac{\mathbb{D}(X)}{\sqrt{\alpha}} < \mathbb{E}(X) < X + \frac{\mathbb{D}(X)}{\sqrt{\alpha}}\right) \geqslant 1 - \alpha.$$

Jeśli $X_1, \ldots, X_n$ jest próba prostą, $\mathbb{E}(X_1) = \mu$, $\mathbb{D}^2(X_1) = \sigma^2$, to wynikający z nierówności Czebyszewa przedział ufności ma postać:

$$\mathbb{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n\alpha}} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n\alpha}}\right) \geqslant 1 - \alpha.$$

Na przykład dla $1 - \alpha = 0.99$ otrzymujemy

$$\mathbb{P}\left(\bar{X} - \frac{10\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{10\sigma}{\sqrt{n}}\right) \geqslant 0.99.$$

W praktyce przedział ten jest (zbyt) duży.

Przedział ufności dla średniej w populacji normalnej o znanej wariancji

- Cecha X ma rozkład $N(\mu, \sigma^2)$, gdzie wariancja σ^2 jest znana.
- Wyznaczymy przedział ufności dla nieznanej wartości parametru μ.
- $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right) \iff Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim N(0, 1).$
- Niech z_α będzie taką liczbą, że

  $$\mathbb{P}(-z_\alpha < Z < z_\alpha) = 1 - \alpha \iff z_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

  Wówczas

  $$\begin{aligned} 1 - \alpha &= \mathbb{P}\left(-z_\alpha < \frac{\bar{X} - \mu}{\sigma}\sqrt{n} < z_\alpha\right) = \\ &= \mathbb{P}\left(-\bar{X} - z_\alpha\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_\alpha\frac{\sigma}{\sqrt{n}}\right) = \\ &= \mathbb{P}\left(\bar{X} - z_\alpha\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_\alpha\frac{\sigma}{\sqrt{n}}\right). \end{aligned}$$

  Otrzymaliśmy zatem

  $$\underline{\theta} = \bar{X} - z_\alpha\frac{\sigma}{\sqrt{n}}, \qquad \overline{\theta} = \bar{X} + z_\alpha\frac{\sigma}{\sqrt{n}}.$$

  Zauważmy, że długość przedziału ufności wynosi tutaj $2z_\alpha\frac{\sigma}{\sqrt{n}}$ i nie zależy od wartości zaobserwowanych w próbie.

Przedział ufności dla średniej w populacji normalnej z nieznaną wariancją

- Cecha X ma rozkład $N(\mu, \sigma^2)$, gdzie wariancja σ^2 jest nieznana.
- Wyznaczymy przedział ufności dla nieznanej wartości parametru μ.
- $t = \dfrac{\bar{X} - \mu}{S}\sqrt{n-1} \sim t_{n-1}$ (gdzie $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$).
- Niech $t_{\alpha,n-1}$ będzie taką liczbą, że

$$\mathbb{P}(-t_{\alpha,n-1} < t < t_{\alpha,n-1}) = 1 - \alpha.$$

Wówczas

$$\begin{aligned} 1 - \alpha &= \mathbb{P}\left(-t_{\alpha,n-1} < \frac{\bar{X} - \mu}{S}\sqrt{n-1} < t_{\alpha,n-1}\right) = \\ &= \mathbb{P}\left(-\bar{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n-1}} < -\mu < -\bar{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n-1}}\right) = \\ &= \mathbb{P}\left(\bar{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n-1}} < \mu < \bar{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n-1}}\right). \end{aligned}$$

Przedział ufności dla średniej w populacji o nieznanym rozkładzie

- Cecha X ma dowolny rozkład, ze znaną wariancją σ^2.
- Wyznaczymy przedział ufności dla nieznanej wartości parametru μ.
- $Z = \dfrac{\bar{X} - \mu}{\sigma}\sqrt{n} \xrightarrow{n\to\infty} N(0, 1)$.
- Zatem, jeśli n jest dostatecznie duże (zwykle wystarcza $n \geqslant 30$), to

$$1 - \alpha = \mathbb{P}\left(\bar{X} - z_\alpha\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_\alpha\frac{\sigma}{\sqrt{n}}\right),$$

gdzie z_α jest taką liczbą, że

$$\mathbb{P}(-z_\alpha < Z < z_\alpha) = 1 - \alpha.$$

- Jeśli σ^2 jest nieznane, to dla dużego n możemy przyjąć $\sigma = S$, otrzymując przedział ufności

$$1 - \alpha = \mathbb{P}\left(\bar{X} - z_\alpha\frac{S}{\sqrt{n}} < \mu < \bar{X} + z_\alpha\frac{S}{\sqrt{n}}\right).$$

Przykład. Zmierzono wytrzymałość 10 losowo wybranych elementów i otrzymano następujące wyniki:

$$383, 284, 339, 340, 305, 386, 387, 335, 344, 346\ [Pa].$$

Przy założeniu, że wytrzymałość tych elementów jest zmienną losową $N(\mu, \sigma^2)$ o nieznanych parametrach μ i σ^2, wyznaczyć na podstawie tej próbki 95% realizację przedziału ufności dla μ. Ponieważ

$$\bar{x} = 344, \quad s_{10}^2 = 986.8, \quad s_{10} = 31.13, \quad t_{0.05,9} = 2.26,$$

więc szukana realizacja przedziału ufności ma postać

$$\left(344 - 2.26 \cdot \frac{31.13}{3}, 344 + 2.26 \cdot \frac{31.13}{3}\right) = (320.5, 367.5).$$

Przedział ufności dla wariancji w populacji normalnej

- Cecha X ma rozkład $N(\mu, \sigma^2)$, z nieznanymi parametrami μ i σ^2.
- Wyznaczymy przedział ufności dla parametru σ^2.
- $\chi^2 = \dfrac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$.
- Wyznaczamy takie liczby $\chi^2_{\alpha/2,n-1}$, $\chi^2_{1-\alpha/2,n-1}$, dla których

  $$\mathbb{P}(\chi^2 \geqslant \chi^2_{\alpha/2,n-1}) = \frac{\alpha}{2}, \qquad \mathbb{P}(\chi^2 \leqslant \chi^2_{1-\alpha/2,n-1}) = \frac{\alpha}{2},$$

  skąd

  $$\mathbb{P}(\chi^2_{1-\alpha/2,n-1} \leqslant \chi^2 \leqslant \chi^2_{\alpha/2,n-1}) = 1 - \alpha.$$

  Wówczas

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

  czyli

  $$\mathbb{P}\left(\frac{nS^2}{\chi^2_{\alpha/2,n-1}} \leqslant \sigma^2 \leqslant \frac{nS^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1 - \alpha.$$

Przykład. W celu zbadania jakości miernika wykonano nim $n = 12$ pomiarów tego samego wzorca. Otrzymano następujące wyniki:

$$275, 273, 279, 267, 276, 272, 271, 269, 270, 265, 268, 277.$$

Przy założeniu, że wyniki pomiarów mają rozkład normalny o nieznanych μ i σ^2, gdzie μ jest prawdziwą wartością wzorca, a σ^2 jest wariancją błędu pomiaru) należy wyznaczyć 90% realizację przedziału ufności dla σ. W wyniku obliczeń otrzymujemy $\bar{x} = 271.8333$, $s = 4.119736$. Ponadto znajdujemy (tablice) $\chi^2_{0.05,11} = 19.67514$, $\chi^2_{0.95,11} = 4.574813$. Podstawiając do powyższego wzoru otrzymujemy przedział ufności dla wariancji σ^2: $(10.35147, 44.51912)$, a stąd dla odchylenia standardowego σ: $(3.217371, 6.672265)$.

Przedział ufności dla wariancji w populacji normalnej, duża próba

- Niech $X_1, \ldots, X_n$ będzie próba prostą z rozkładu $N(\mu, \sigma^2)$ gdzie μ i σ^2 są nieznane, natomiast $n > 30$.
- Możemy skorzystać z faktu, iż statystyka S_n ma asymptotyczny rozkład $N(\sigma, \sigma/\sqrt{2n})$, więc

  $$Z = \frac{S_n - \sigma}{\sigma}\sqrt{2n} \xrightarrow{n\to\infty} N(0, 1).$$

- Zatem dla $z_\alpha = \Phi^{-1}(1 - \alpha/2)$ mamy

  $$\mathbb{P}\left(-z_\alpha < \frac{S_n - \sigma}{\sigma}\sqrt{2n} < z_\alpha\right) \approx 1 - \alpha,$$

  więc przedział ufności dla σ na poziomie ufności $1 - \alpha$ ma postać

  $$\mathbb{P}\left(\frac{S_n}{1 + \frac{z_\alpha}{\sqrt{2n}}} < \sigma < \frac{S_n}{1 - \frac{z_\alpha}{\sqrt{2n}}}\right) = 1 - \alpha,$$

  lub w przybliżeniu

  $$\mathbb{P}\left(S_n\left(1 - \tfrac{z_\alpha}{\sqrt{2n}}\right) < \sigma < S_n\left(1 + \tfrac{z_\alpha}{\sqrt{2n}}\right)\right) = 1 - \alpha.$$

Przedział ufności dla frakcji (prosty)

- Cecha ma rozkład zero-jedynkowy z nieznanym parametrem p.
- Niech X oznacza liczbę sukcesów w próbie n-elementowej.
- Jeśli n jest dostatecznie duże oraz $0.04 \leqslant p \leqslant 0.96$, to w przybliżeniu

$$W = \frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Po zestandaryzowaniu i zastąpieniu w mianowniku p przez jego estymator W otrzymujemy:

$$Z = \frac{W-p}{\sqrt{\frac{W(1-W)}{n}}} \sim N(0,1).$$

Zatem jeśli z_α jest taką liczbą, że $\mathbb{P}(-z_\alpha < Z < z_\alpha) = 1-\alpha$, to

$$\begin{aligned} 1-\alpha = \mathbb{P}\left(-z_\alpha < \frac{W-p}{\sqrt{\frac{W(1-W)}{n}}} < z_\alpha\right) = \\ = \mathbb{P}\left(W - z_\alpha\sqrt{\frac{W(1-W)}{n}} < p < W + z_\alpha\sqrt{\frac{W(1-W)}{n}}\right). \end{aligned}$$

Niestety w przypadku p bliskich 0 lub 1, końce tego przedziału ufności mogą wychodzić poza przedział $[0,1]$.

Przedział ufności dla frakcji (lepszy)

- Jak poprzednio, cecha ma rozkład zero-jedynkowy z nieznanym parametrem p, zaś X oznacza liczbę sukcesów w próbie n-elementowej, gdzie n jest wystarczająco duże ($n > 100$).
- Mamy

$$\frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right) \iff \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{X-np}{\sqrt{np(1-p)}} \sim N(0,1).$$

skąd

$$1-\alpha = \mathbb{P}\left(|Z| < z_\alpha\right) = \mathbb{P}\left(\left|\frac{X-np}{\sqrt{np(1-p)}}\right| < z_\alpha\right).$$

Rozwiązując nierówności w ostatnim nawiasie względem p, uzyskamy szukany przedział ufności

$$A(B-C) < p < A(B+C),$$

gdzie

$$A = \frac{1}{n+z_\alpha^2}, \quad B = X + \frac{z_\alpha^2}{2}, \quad C = z_\alpha \cdot \sqrt{\frac{X(n-X)}{n} + \frac{z_\alpha^2}{4}}.$$

Problem minimalnej liczebności próby

- $d^* = \dfrac{\overline{\theta} - \underline{\theta}}{2}$ – maksymalny błąd szacunku.
- Dla ustalonej wartości d dobieramy liczebność próby, aby $d^* \leqslant d$.

- W przypadku średniej w populacji normalnej ze znaną wariancją mamy $d^* = z_\alpha \frac{\sigma}{\sqrt{n}}$, więc

$$z_\alpha \frac{\sigma}{\sqrt{n}} \leqslant d \iff z_\alpha \frac{\sigma}{d} \leqslant \sqrt{n} \iff n \geqslant \frac{z_\alpha^2 \sigma^2}{d^2}.$$

- W przypadku frakcji $d^* = z_\alpha \sqrt{\frac{p(1-p)}{n}}$, zatem

$$z_\alpha \sqrt{\frac{p(1-p)}{n}} \leqslant d \iff z_\alpha \frac{\sqrt{p(1-p)}}{d} \leqslant \sqrt{n} \iff n \geqslant \frac{z_\alpha^2 p(1-p)}{d^2}.$$

- Jeśli nie mamy informacji o wielkości p, to zawsze możemy szacować z góry:

$$p(1-p) \leqslant \frac{1}{4}.$$

Przykład. Przypuśćmy, że w badaniach (poparcia dla kandydata w wyborach) interesuje nas liczność próby wystarczająca do wyznaczenia przedziału ufności na poziomie ufności 0.9, którego dopuszczalna długość nie przekracza $5\% = 0.05$. Otrzymujemy warunek

$$\frac{z_{0.95}}{\sqrt{4n}} \leqslant \frac{0.05}{2} \iff n \geqslant \frac{z_{0.95}^2}{4 \cdot 0.025^2} = \frac{1.6449^2}{4 \cdot 0.025^2} = 1083.$$

Zazwyczaj po przeprowadzeniu badania długość przedziału ufności będzie mniejsza. Na przykład, gdy $n = 1083$, $X = 345$, to $W = \frac{345}{1083} = 0.3186$, $\sqrt{\frac{W(1-W)}{n}} = 0.01416$, a realizacja przedziału ufności dla p ma postać:

$$(0.295, 0.342) = (29.5\%, 34.2\%).$$