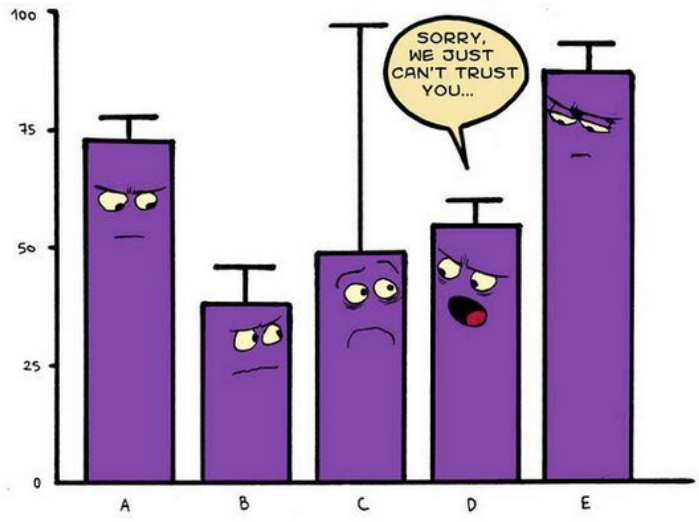


5. Symulacje. Estymacja punktowa.

5.1 Symulacje

5.1.1 Uniwersalność rozkładu jednorodnego

W komputerze generujemy (symulujemy) zdarzenia losowe poprzez wykorzystanie tzw. generatorów liczb pseudolosowych. Taki generator to nic innego jak bardzo skomplikowany wzór matematyczny nakładany na jedną lub kilka zmiennych przechowywanych w generatorze. Wzór ten jest deterministyczny tzn. gdy zainicjujemy zmienne generatora w taki sam sposób to zawsze otrzymamy ten sam ciąg liczb pseudolosowych. Z tego powodu ważne jest inicjalizowanie generatora różnymi wartościami początkowymi np. liczbą pobraną z zegara systemowego, choć w niektórych sytuacjach np. przy debugowaniu powtarzalność generacji liczb pseudolosowych dla tej samej wartości początkowej może być bardzo użyteczna.

Jednakże, taki generator zwykle generuje liczby pseudolosowe o rozkładzie jednorodnym z przedziału $[0, 1)$. Co zrobić jeśli interesuje nas generowanie liczb z dowolnego rozkładu? Jeżeli interesuje nas inny rozkład jednorodny możemy to zwykle osiągnąć poprzez odpowiednie wymnożenie i dodanie/odjęcie jakieś wartości od tego rozkładu, ale co z rozkładem normalnym, Poissona itd.? Przypomnijmy sobie rozkład jednorodny, a potem poznajmy jego bardzo użyteczną własność, która pozwoli nam na rozwiązanie tego problemu.

Definicja 5.1 — Rozkład jednorodny. Rozkład jednorodny (jednostajny, równomierny, prostokątny) to ciągły rozkład prawdopodobieństwa, dla którego gęstość prawdopodobieństwa w przedziale od a do b, jest stała i różna od zera, a poza nim równa zeru.

Problem 5.1 — Własności rozkładu jednostajnego. Odpowiedz na poniższe pytania:

- Jaką wartość przyjmuje funkcja gęstości prawdopodobieństwa w przedziale od a do b?

- Ile wynosi wartość oczekiwana?
- Ile wynosi wariancja?

Problem 5.2 W jaki sposób możesz wygenerować liczby losowe z rozkładu Bernoulliego z prawdopodobieństwem sukcesu p?

Twierdzenie 5.1 — Uniwersalność rozkładu jednorodnego. Niech F będzie ciągłą, silnie rosnącą [a] dystrybuantą wtedy $X = F^{-1}(U) \sim F$ jeżeli $U \sim Uniform(0,1)$.

Także jeżeli $X \sim F$ to wtedy $F(X) \sim Uniform(0,1)$.

[a] Zależy nam na tym, aby F^{-1} była dobrze określona. Jeżeli zdefiniujemy $F^{-1}(x) = \inf\{y : F(y) \geq x\}$ to możemy pominąć to wymaganie.

5.1.2 Metody Monte Carlo

Metoda Monte Carlo jest stosowana do modelowania matematycznego procesów zbyt złożonych (obliczania całek, łańcuchów procesów statystycznych), aby można było przewidzieć ich wyniki za pomocą podejścia analitycznego. Istotną rolę w metodzie Monte Carlo odgrywa losowanie (wybór przypadkowy) wielkości charakteryzujących proces, przy czym losowanie dokonywane jest zgodnie z rozkładem, który musi być znany. Metoda została opracowana i pierwszy raz zastosowana przez Stanisława Ulama [1], polsko-amerykańskiego matematyka.

Problem 5.3 W jaki sposób możemy uzyskać oszacowanie liczby π używając generatora liczb pseudolosowych?

Problem 5.4 W jaki sposób możemy przybliżyć $F(x) = P(X \leq x)$ umiejąc wylosować liczby z danego rozkładu?

Ćwiczenie 5.1 — Metody Monte Carlo. Otwórz arkusz kalkulacyjny dostępny pod następującym linkiem: `https://ophelia.cs.put.poznan.pl/webdav/ad/students/03/cw-2.xls` i rozwiąż ćwiczenie. ■

5.2 Estymacja punktowa

Na poprzednich laboratoriach poznaliśmy statystykę matematyczną, która ściśle łączy statystykę z rachunkiem prawdopodobieństwa. Zakłada ona, że analizowane dane powstały z jakiegoś ukrytego rozkładu prawdopodobieństwa. W niektórych sytuacjach o tym ukrytym rozkładzie nie wiemy zupełnie nic, jednak dość często możemy go przybliżyć którymś ze znanych rozkładów prawdopodobieństwa np. rozkładem normalnym. W takim przypadku głównym naszym zadaniem jest oszacowanie parametrów takiego rozkładu: w przypadku rozkładu normalnego jest to μ i σ, a w przypadku rozkładu Bernoullego jest to prawdopodobieństwo sukcesu p. Aby oszacować taki parametr obliczamy pewną funkcję na otrzymanej z populacji próbie losowej, której wynik będziemy traktować jako estymację tego parametru.

Problem 5.5 Jakie własności musi spełniać próba losowa prosta?

Definicja 5.2 — Statystyka. Statystyką (ang. *statistic*) nazywa się liczbową charakterystykę próby $T(X_1, X_2, \ldots)$. Statystyka jest więc zmienną losową.

Definicja 5.3 — Estymator. Estymatorem $\hat{\theta}$ nazywamy statystykę służącą do oszacowania nieznanego parametru populacji θ.

Załóżmy, że chcemy wyestymować wartość oczekiwaną badanej cechy statystycznej. Jak możemy to zrobić? Prawdopodobnie pierwszym estymatorem, który przychodzi ci do głowy jest po prostu policzenie średniej arytmetycznej dla tych danych ($\bar{x}$). Jednak nie jest to jedyna możliwa odpowiedź!

Po pierwsze twój leniwy kolega może stwierdzić, że prawdziwa wartość oczekiwana tak czy tak nie jest znana, więc dlaczego by nie użyć wartości jednego, losowo wybranego elementu z próby jako estymację wartości średniej $\hat{\theta}_{random} = RandomChoice(X_1, X_2, ...)$?

Ktoś inny może stwierdzić, że wartość średnia w próbce może być zaniżona w stosunku do prawdziwej i należałoby ją przemnożyć przez $\sqrt[n]{n}$. Mielibyśmy wtedy $\hat{\theta}_{zwiekszone} = \sqrt[n]{n}\frac{1}{n}\sum_{i=1}^{n} x_i$. Na przykład dla małej, 20-elementowej próby zwiększylibyśmy średnią o ok. 16% ($\sqrt[20]{20} \approx 1,1615$), a dla dużej próby o 1000 elementów nie zwiększylibyśmy jej praktycznie wcale ($\sqrt[1000]{1000} \approx 1,0069$).

Z kolei jakiś inżynier pracujący na dużych danych, może stwierdzić, że policzenie średniej wartości trwa bardzo długo (złożoność jest liniowa, ale n jest bardzo duże). Z tego powodu może zaproponować szybszy sposób liczenia średniej poprzez policzenie jej dla co 10 obserwacji $\hat{\theta}_{co10} = \frac{1}{n/10}\sum_{i=1}^{n/10} x_{10i}$. Dodatkowo takie postępowanie można motywować faktem, że w praktyce czasami w liczeniu średniej pomijamy trochę obserwacji (np. średnia trymowana).

Takich pomysłów na estymator średniej może być znaczenie więcej, ale nie podajemy ich tutaj ze względu na łatwość dalszego postępowania matematycznego. Zasygnalizujmy jednak, że wcale nie musi być to modyfikacja standardowego wzoru na średnią arytmetyczną, a może to być zupełnie inna funkcja np. dla rozkładów symetrycznych estymatorem średniej (i to całkiem niezłym) może być mediana.

Każdy z tych pomysłów ma jakąś stojącą za nim logikę i ciężko od razu powiedzieć, który z tych pomysłów jest najlepszy. Problem jest jeszcze większy, gdybyśmy chcieli estymować jakieś inne, bardziej skomplikowane parametry. Znamy np. dwa wzory na wariancję w próbie i w populacji - którego z nich powinno się użyć? Który z nich jest lepszy? Poszukajmy więc pewnych dobrych własności, które chcielibyśmy aby spełniała nasza estymacja.

Problem 5.6 Jakie powinny być własności dobrego estymatora?

Jedną z cech dobrego estymatora jest to, żeby wraz ze zwiększającą się próbą otrzymujemy coraz to lepsze oszacowania szukanego parametru. W celu otrzymania lepszego oszacowania potrzebujemy pobrać większą próbkę (co by to było gdyby się okazało że na mniejszej i tańszej próbce nasze estymacje są lepsze?). Rozsądne wydaje się też wymaganie, że przy rozmiarze próby rosnącym do nieskończoności ryzyko błędu było coraz to niższe, a w końcu zerowe. Estymatory posiadające taką własność nazywamy estymatorami zgodnymi.

Definicja 5.4 — Estymator zgodny. Estymator $\hat{\theta}_n$ posiadający własność

$$\forall_{\varepsilon>0} \quad \lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$$

nazywamy estymatorem zgodnym (ang. *consistent estimator*).

Które z naszych estymatorów $\hat{\theta}_{co10}$, $\hat{\theta}_{random}$, $\hat{\theta}_{zwiekszone}$, $\bar{x}$ mają taką własność? Sprawdźmy ją najpierw dla zwykłej średniej arytmetycznej.

■ **Przykład 5.1** Udowodnij, że $\bar{X}_n$ na próbie prostej jest estymatorem zgodnym wartości oczekiwanej[1]. Ponieważ mamy próbę prostą, zmienne X_i mają taki sam rozkład czyli tą samą wartość oczekiwaną i wariancję. Z tego powodu dalej będziemy używać oznaczeń $\mathbb{E}[X_i] = \mu$ i $\mathbb{D}^2[X_i] = \sigma^2$.

Używając nierówności Czebyszewa mamy

$$P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\mathbb{D}^2(\bar{X}_n)}{\varepsilon^2}$$

Jaka jest wariancja średniej z próby?

$$\begin{aligned}
\mathbb{D}^2(\bar{X}_n) &= \mathbb{D}^2(\frac{1}{n}\sum_{i=1}^{n} X_i) \\
&= \mathbb{D}^2(\frac{1}{n}X_1 + \frac{1}{n}X_2 + ... + \frac{1}{n}X_n) \\
&= \mathbb{D}^2(\frac{1}{n}X_1) + \mathbb{D}^2(\frac{1}{n}X_2) + ... + \mathbb{D}^2(\frac{1}{n}X_n) \text{ [mamy do czynienia z próbą prostą, a więc} \\
&\text{zmienne są niezależne czyli } \mathbb{D}^2(X+Y) = \mathbb{D}^2X + \mathbb{D}^2Y] \\
&= \frac{1}{n^2}\mathbb{D}^2(X_1) + \frac{1}{n^2}\mathbb{D}^2(X_2) + ... + \frac{1}{n^2}\mathbb{D}^2(X_n) \text{ [ze wzoru } \mathbb{D}^2(aX+b) = a^2\mathbb{D}^2(X)] \\
&= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + ... + \frac{1}{n^2}\sigma^2 \text{ [zmienne mają jednakowy rozkład } \mathbb{D}^2(X_i) = \sigma^2] \\
&= \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}$$

Wzór wyprowadzony powyżej będzie przydatny na tych i na kolejnych laboratoriach z tego powodu warto go zapamiętać. Zauważ, że wariancja tego estymatora jest mniejsza od wariancji badanej zmiennej losowej oraz, że jest tym niższa im większa jest próba.

Wracając do nierówności Czebyszewa:

$$P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\mathbb{D}^2(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

$$\lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0$$

Podsumowując:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = \lim_{n\to\infty} 1 - P(|\bar{X}_n - \mu| > \varepsilon) = 1$$

czyli $\bar{X}_n$ jest estymatorem zgodnym wartości oczekiwanej. ■

[1] Jeżeli ktoś pamięta prawo wielkich liczb to z pewnością zauważył, że zgodność tego estymatora wypływa bezpośrednio z tego prawa. Tutaj prezentuję rozumowanie wykorzystujące znajomość poprzednich laboratoriów (nierówność Czebyszewa)

Udowodniliśmy, że średnia arytmetyczna jest estymatorem zgodnym wartości oczekiwanej. Dla pozostałych estymatorów nie będziemy przeprowadzać dowodu, a tylko sprawdzimy to intuicyjnie. Po pierwsze $\hat{\theta}_{co10}$ również będzie zgodny, bo przy n dążącym do nieskończoności fakt nie wzięcia pod uwagę pewnej części wartości nie będzie miał znaczenia (skoro i tak dążymy do nieskończoności). Co interesujące $\hat{\theta}_{zwiekszone}$ również będzie zgodny, bo $\sqrt[n]{n}$ w granicy wynosi 1, a mnożenie przez 1 nie zmieni wyniku. Jedynym estymatorem, który nie posiada tej własności jest $\hat{\theta}_{random}$, gdyż będzie on tak samo dobry (lub jak kto woli, tak samo zły) dla każdej wartości n. Nie ma większego związku pomiędzy jego jakością estymacji, a wielkością próbki.

Jaką jeszcze własność mógłby mieć estymator? Warunek zgodności mówi o prawidłowym oszacowaniu szukanego parametru przy rosnącym w nieskończoność rozmiarze próby. Próba o nieskończonym rozmiarze z oczywistych względów nie jest możliwa w praktyce, a często jesteśmy zmuszeni pracować z małymi próbami. Szczególnie wtedy gdy nagromadzenie danych jest kosztowne. W związku z tym chcielibyśmy, aby wartość estymatora trafiała mniej więcej w prawdziwą wartość szacowanego parametru bez względu na wielkość próby. Jak możemy taki warunek ująć w zapisie matematycznym? Poprzez zapewnienie, że wartość oczekiwana estymatora będzie równa prawdziwej wartości szacowanego parametru.

Definicja 5.5 — Estymator nieobciążony. Estymator $\hat{\theta}_n$ nazywamy nieobciążonym jeżeli

$$\underset{n}{\forall} \quad \mathbb{E}[\hat{\theta}_n] = \theta$$

W przeciwnym wypadku estymator nazywamy obciążonym z obciążeniem wynoszącym $B(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$.

Problem 5.7 Sprawdź czy $\bar{x}$ jest estymatorem nieobciążonym.

Zauważ, że estymator $\hat{\theta}_{co10}$ także będzie nieobciążony – jest to tak naprawdę estymator $\bar{x}$ policzony na 10-krotnie mniejszej próbie, więc jego wartość oczekiwana $\hat{\theta}_{zwiekszone}$ jest także równa wartości oczekiwanej szukanego parametru. Natomiast estymator $\hat{\theta}_{zwiekszone}$ nie będzie miał już tej własności, ponieważ jego wartość oczekiwana będzie wynosiła $\sqrt[n]{n}\mu$.

Problem 5.8 Czy estymator może nie być zgodny, ale być nieobciążony?

Na naszym polu bitwy zostały dwa estymatory $\bar{x}$ i $\hat{\theta}_{co10}$. Który z nich jest lepszy? Cóż oba niezależnie od rozmiaru próby średnio trafiają w szukaną wartość parametru populacji (czyli w naszym przypadku w wartość oczekiwaną), natomiast dodatkowo chcielibyśmy, aby nasz szukany estymator trafiał w tę wartość jak najdokładniej tj. z najmniejszą możliwą wariancją. Taki estymator nazywamy estymatorem efektywnym. Czasami tłumaczy się go jako estymator najefektywniejszy i jest to chyba bardziej intuicyjna nazwa, ponieważ podkreśla ona że nie istnieje estymator efektywniejszy (o mniejszej wariancji) od estymatora efektywnego.

Definicja 5.6 — Estymator efektywny. Estymator nieobciążony $\hat{\theta}_n$ nazywamy efektywnym (lub najefektywniejszym) jeżeli ma on najmniejszą wariancję spośród wszystkich estymatorów nieobciążonych. Taki estymator nie zawsze istnieje.

Problem 5.9 Który z rozważanych estymatorów nieobciążonych $\bar{x}$, $\hat{\theta}_{co10}$ jest efektywniejszy?

Poznaliśmy estymatory dla wartości oczekiwanej i pokazaliśmy że $\bar{x}$ jest estymatorem zgodnym i nieobciążonym. Nie będziemy tutaj przeprowadzać dowodu, ale można dodatkowo pokazać, że estymator ten jest estymatorem efektywnym dla rozkładu normalnego. Jakie są jednak estymatory dla wariancji? Z poprzednich zajęć możemy zaproponować dwa: $S^{2*} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ oraz $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Oba te estymatory są estymatorami zgodnymi, ale czy nieobciążonymi?

■ **Przykład 5.2** Sprawdźmy czy S^{2*} jest estymatorem nieobciążonym?

$$\begin{aligned} S^{2*} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{1}{n}\sum_{i=1}^{n}2X_i\bar{X} + \frac{1}{n}\sum_{i=1}^{n}\bar{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\bar{X}\frac{1}{n}\sum_{i=1}^{n}X_i + \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\bar{X}^2 + \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2 \end{aligned}$$

stąd

$$\mathbb{E}[S^{2*}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[\bar{X}^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[\bar{X}^2]$$

Wiemy ile wynosi wariancja średniej arytmetycznej ($\mathbb{D}^2\bar{X} = \frac{\mathbb{D}^2X}{n} = \frac{\sigma^2}{n}$), więc chcielibyśmy skorzystać z tej wiedzy w naszym wyprowadzeniu. Spróbujmy wykorzystać wzór na wariancję $\mathbb{D}^2X = \mathbb{E}X^2 - (\mathbb{E}X)^2$ dla wariancji średniej arytmetycznej. Z tego powodu dodajemy i odejmujemy $(\mathbb{E}\bar{X})^2 = (\mathbb{E}X)^2 = \mu^2$.

$$\begin{aligned} \mathbb{E}[S^{2*}] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - (\mathbb{E}[\bar{X}^2] - \mu^2 + \mu^2) \text{ /*stosujemy wzór na wariancję*/} \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2 - \mu^2 + \mu^2] - (\frac{\sigma^2}{n} + \mu^2) \text{ /*jeszcze raz trik z odjęciem i dodaniem } \mu^2 \text{ */} \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}[X_i^2] - \mu^2) + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mu^2] - (\frac{\sigma^2}{n} + \mu^2) \text{ /* } \mathbb{E}[X_i^2] - \mu^2 \text{ to nasz wzór na wariancję! */} \\ &= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 \\ &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

Jest to więc estymator obciążony, a mnożąc go przez $\frac{n}{n-1}$ otrzymamy estymator nieobciążony:

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

Jeśli powyższe wyprowadzenie było dla Ciebie za szybkie poniżej prezentuje drugi, dłuższy, ale prostszy w analizie krok po kroku sposób wyprowadzania powyższego wyniku.

$$
\begin{aligned}
S^{2*} &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu+\mu-\bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}((X_i-\mu)+(\mu-\bar{X}))^2 \text{ [wykorzystajmy wzór skróconego mnożenia]} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left((X_i-\mu)^2+2(X_i-\mu)(\mu-\bar{X})+(\mu-\bar{X})^2\right) \text{ [przesuńmy sumę do środka]} \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+\frac{1}{n}\sum_{i=1}^{n}2(X_i-\mu)(\mu-\bar{X})+\frac{1}{n}\sum_{i=1}^{n}(\mu-\bar{X})^2 \text{ [stałe na zewnątrz]} \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+2(\mu-\bar{X})\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)+\frac{1}{n}n(\mu-\bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+2(\mu-\bar{X})\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)+(\mu-\bar{X})^2 \text{ [środkowa suma do środka]} \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+2(\mu-\bar{X})(\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}\mu)+(\mu-\bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+2(\mu-\bar{X})(\bar{X}-\mu)+(\mu-\bar{X})^2 \text{ [nawias w środku razy -1]} \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-2(\mu-\bar{X})(\mu-\bar{X})+(\mu-\bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\mu-\bar{X})^2
\end{aligned}
$$

Przejdźmy do obliczenia wartości oczekiwanej:

$$
\begin{aligned}
\mathbb{E}[S^{2*}] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\mu-\bar{X})^2] \text{ [liniowość oczekiwań: } \mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y]] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i-\mu)^2]-\mathbb{E}[(\mu-\bar{X})^2] \text{ [zauważ że } \mathbb{E}[(X_i-\mu)^2 \text{ to wariancja]} \\
&= \frac{1}{n}n\sigma^2-\mathbb{E}[(\mu-\bar{X})^2] \text{ [po prawej też jest prawie wzór na wariancję – pomnóż przez -1]} \\
&= \sigma^2-\mathbb{E}[(\bar{X}-\mu)^2] \text{ [po prawej mamy teraz wariancję średniej arytmetycznej]} \\
&= \sigma^2-\frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2
\end{aligned}
$$

■

Ćwiczenie 5.2 — Zgodność estymatorów. Otwórz arkusz kalkulacyjny dostępny pod następującym linkiem: `https://ophelia.cs.put.poznan.pl/webdav/ad/students/04/cw-3.xls` i rozwiąż ćwiczenie. ■

Ćwiczenie 5.3 — Obciążenie estymatorów. Otwórz arkusz kalkulacyjny dostępny pod następującym linkiem: `https://ophelia.cs.put.poznan.pl/webdav/ad/students/04/cw-1.xls` i rozwiąż ćwiczenie. ■

Problem 5.10 Masz daną funkcję gęstości prawdopodobieństwa z kilkoma parametrami. W jaki sposób mógłbyś wyznaczyć estymatory tych parametrów?

5.3 * Jak wyznaczać estymatory?

Podczas tłumaczenia zagadnienia na tablicy poznaliśmy kilka dziwnych estymatorów (dla średniej) i wiemy, że jest tutaj duże pole do popisu dla naszej wyobraźni. Wiemy też, że chcielibyśmy aby nasz estymator miał jak najwięcej „fajnych" własności np. zgodność. Jak jednak możemy stworzyć/wymyślić taki estymator? Możemy oczywiście próbować atakować problem metodą prób i błędów: wymyślać estymator i udowadniać jego własności (lub raczej nie być w stanie ich udowodnić, bo wymyślane estymatory tych własności po prostu nie będą miały)[2]

Zastanówmy się w jaki sposób moglibyśmy takie estymatory tworzyć. Rozważmy prosty przykład: otrzymujemy dane i chcemy zamodelować je rozkładem normalnym. Musimy więc oszacować parametry tego rozkładu μ i σ. Rysunek 5.1 (po lewej) przedstawia taką przykładową sytuację. Który z rozkładów byś wybrał? Rozkład, który jest narysowany na niebiesko czy rozkład czerwony? Prawdopodobnie wybrałbyś rozkład niebieski. Dlaczego? Tu odpowiedzią może być np. bo rozkład niebieski ma średnią bliższą średniej danych[3]. Po prawej stronie rysuneku 5.1 masz inne dane, które próbujesz zamodelować rozkładem χ^2 (spokojnie, nie musisz tego rozkładu znać – jeszcze:). Prawdopodobnie znów wybrałbyś rozkład niebieski, choć z rysunku ciężko ocenić czy średnia tego rozkładu jest bliższa średniej danych (zauważ obserwacje odstające!). Dlaczego? Bo ten rozkład (podobnie jak niebieski rozkład normalny) lepiej *pasuje* do danych tzn. rozkłady te czynią otrzymanie/wylosowanie takich danych bardziej prawdopodobne.

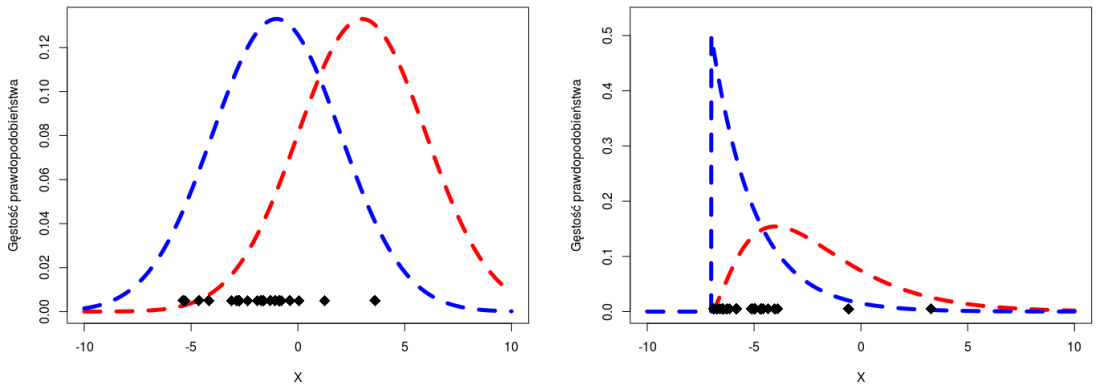
5.3.1 Estymacja największej wiarygodności

Sposób który właśnie intuicyjnie odkryliśmy to popularna technika tworzenia estymatorów w statystyce częstotliwościowej nazywana estymacją maksymalnej wiarygodności (ang. *maximum likelihood estimation, MLE*). Idea tego sposobu jest następująca: wybieramy parametry rozkładu tak, aby jak najlepiej pasował on do naszych danych. Co to znaczy, że rozkład najlepiej pasuje do naszych danych? To znaczy, że wśród wszystkich możliwych rozkładów opisanych tym parametrem, wybrany rozkład przypisuje największe możliwe prawdopodobieństwo do uzyskania naszego zbioru danych. Krótko mówiąc: szukamy rozkładu, który czyni nasze dane najbardziej prawdopodobnymi[4].

[2] Możemy mieć problemy już na etapie „wymyślania estymatorów". Problem estymacji, który wcześniej rozważaliśmy dotyczył (bajecznie prostej) średniej arytmetycznej, ale co jeśli estymujemy parametry 1000-wymiarowego rozkładu normalnego?

[3] De facto jest metoda tworzenia estymatorów, która tworzy je w ten sposób: metoda momentów. Jest ona jednak słabsza niż poznawana przez nas metoda największej wiarygodności.

[4] Przypomnijmy: zakładamy, że populacja ma pewien rozkład opisany parametrami θ (np. populacja jest modelowana rozkładem normalnym z parametrami μ, σ). Z tego rozkładu (populacji) losujemy próbkę n elementową. Uzyskanie takiej, a nie innej próbki ma pewne prawdopodobieństwo. I to właśnie prawdopodobieństwo będziemy w MLE maksymalizować



Rysunek 5.1: Przykłady otrzymanych danych (czarne romby) oraz dwie propozycje modelowania ich rozkładami z tej samej rodziny (rozkłady normalne i rozkłady χ^2). Estymowanie parametrów tych rozkładów jakie powinno nam dać parametry? Parametry rozkładów niebieskich czy czerwonych?

Jak policzyć prawdopodobieństwo otrzymania konkretnych danych? Jak zwykle zakładamy, że nasza próba jest próbą losową prostą, a więc że każda kolejna obserwacja jest niezależna. Możemy więc pomnożyć prawdopodobieństwo każdej kolejnej obserwacji, aby uzyskać prawdopodobieństwo uzyskania całej próbki (czyli serii obserwacji). Zapisujemy to wzorem:

$$L(\theta) = \prod_{i=1}^{n} P(x_i|\theta)$$

gdzie $P(x_i|\theta)$ oznacza prawdopodobieństwo uzyskania $X = x_i$ z rozkładu opisanego parametrem θ). Wartość powyższej funkcji wskazuje nam jak bardzo prawdopodobne jest zaobserwowanie wartości $x_1, ..., x_n$ przy parametrach θ.

■ **Przykład 5.3** Rozważmy prosty przykład. Rzucamy pięć razy monetą i otrzymujemy 4 orły i 1 reszkę. Zakładamy rozkład dwumianowy, który ma tylko jeden parameter p (a więc nasze θ to po prostu p). Porównajmy ze sobą 2 możliwe rozkłady: $B(0.5)$ oraz $B(0.8)$.

Funkcja $L(p)$ dla naszych danych i $p = 0.5$ wynosi:

$$L(p) = \prod_{i=1}^{n} P(x_i|p) = \left(\frac{1}{2}\right)^4 \cdot \frac{1}{2} = 0,03125$$

Funkcja $L(p)$ dla naszych danych i $p = 0.8$ wynosi:

$$L(p) = \prod_{i=1}^{n} P(x_i|p) = \left(\frac{4}{5}\right)^4 \cdot \frac{1}{5} = 0,08192$$

Widzimy więc, że model drugi lepiej pasuje do danych niż model pierwszy. ■

Estymacja maksymalnej wiarygodności to taka, która zwraca parametry θ które maksymalizują funkcję $L(\theta)$. Prawdopodobnie pamiętasz z analizy matematycznej (lub z liceum) jak znajdywać maksymalną wartość funkcji (słowo-klucz: pochodne). Często dużo prościej (na kartce) i dużo bardziej stabilnie numerycznie (na komputerze) jest maksymalizować logarytm wiarygodności[5] (log-likelihood). Dlaczego? Nie trudno się domyślić: logarytm z mnożeń (zobacz jak jest skonstruowana funkcja L!) to suma logarytmów. Z kolei pochodna sumy to suma pochodnych, co znakomicie upraszcza nam obliczenia[6].

$$\ell(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln(P(x_i|\theta))$$

■ **Przykład 5.4** Zebraliśmy dane pochodzące z populacji modelowanej rozkładem wykładniczym[7]. Rozkład ten przyjmuje formę:

$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Wyznacz estymator największej wiarygodności dla parametrów tego rozkładu (rozkład ten ma jeden parametr λ).

W przypadku rozkładu wykładniczego funkcja wiarygodności przyjmuje postać[8]:

$$\ell(\lambda) = \sum_{i=1}^{n} \ln\left(P(x_i|\lambda)\right) = \sum_{i=1}^{n} \ln\left(\lambda e^{-\lambda x_i}\right) = \sum_{i=1}^{n} (\ln\lambda - \lambda x_i) = n \cdot \ln\lambda - \lambda \sum_{i=1}^{n} x_i$$

Policzmy pochodną[9]:

$$\frac{d\ell}{d\lambda} = \frac{d[n \cdot \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i]}{d\lambda} = n\frac{d\ln(\lambda)}{d\lambda} - \sum_{i=1}^{n} x_i \frac{d\lambda}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

Przyrównajmy ją do zera:

$$\frac{d\ell}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\lambda = \frac{n}{\sum_{i=1}^{n} x_i},$$

Uzyskaliśmy estymator największej wiarygodności dla rozkładu wykładniczego.

! Formalnie powinniśmy sprawdzić jeszcze czy jest to maksimum licząc drugą pochodną: $\frac{d\ell}{d^2\lambda} = -\frac{n}{\lambda^2}$ co jest zawsze mniejsze od 0 (bo liczba obserwacji $n > 0$).

■

Podsumujmy więc nasze rozważania podajac kilka definicji:

[5] Logarytm wiarygodności osiąga maksimum dla tego samego parametru co wiarygodność, ponieważ logarytm jest ściśle rosnący dla dziedziny $[0,1]$.

[6] Zamiast wielokrotnie mnożyć, a następnie liczyć pochodna z tego skomplikowanego tworu, możemy policzyć pochodną z pojedynczego $\ln P(x_i|\lambda)$ a następnie je zsumować!

[7] Rozkład wykładniczy (ang. *exponential distribution*) obserwujemy kiedy opisujemy długość czasu pomiędzy kolejnymi zdarzeniami w homogenicznym procesie Poisson'a.

[8] Podstawowe własności: $\ln(ab) = \ln a + \ln b$ oraz $\ln(e^a) = a$

[9] Pamiętaj że $(\ln x)' = \frac{1}{x}$ oraz $(f(x) + g(x))' = f'(x) + g'(x)$

Definicja 5.7 — Funkcja wiarygodności. Funkcja wiarygodności jest zdefiniowana jako

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

gdzie $f(x_i|\lambda)$ oznacza odpowiednio funkcję gęstości prawdopodobieństwa (zmienne ciągłe) lub funkcję masy prawdopodobieństwa (zmienne dyskretne) dla rozkładu opisanego parametrem λ).

Definicja 5.8 — Estymator największej wiarygodności. Estymator największej wiarygodności (MLE) oznaczany $\hat{\theta}_{MLE}$ przyjmuje wartość θ, która maksymalizuje funkcję wiarygodności $L(\theta)$.

Tyle się namęczyliśmy, ale skąd wiemy czy MLE są dobrymi estymatorami? Jakie własności mają estymatory MLE? Otóż okazuje się, że pod pewnymi warunkami[10] MLE mają wiele użytecznych własności:

- MLE są zgodne
- MLE są asymptotycznie nieobciążone - estymator MLE dla dużych prób jest estymatorem nieobciążonym
- MLE jest asymptotycznie efektywny - estymator MLE ma najmniejszą możliwą wariancję spośród wszystkich dobrze zachowujących się estymatorów dla dużych prób losowych.
- MLE są asymptotycznie normalne - estymator MLE dąży do rozkładu normalnego dla dużych prób losowych.
- i wiele innych ;)

Jeśli pomnożymy $L(\theta)$ przez jakąkolwiek stałą dodatnią c, nie zmienimy ostatecznego estymatora MLE. Z tego powodu często dla uproszczenia obliczeń przy optymalizacji pomijamy stałe.

5.3.2 Czy estymator nieobciążony jest zawsze lepszy od obciążonego?

Oczywiście, ktoś sceptycznie nastawiony może powiedzieć, że MLE dla małych prób są jedynie zgodne, w dodatku mogą być obciążone! Chociaż używanie estymatora nieobciążonego jest dobrym pomysłem, w nietórych sytuacjach może się okazać że estymator (troszczkę) obciążony będzie w danym zastosowaniu lepszy, bo np. będzie miał dużo mniejszą wariancję.

Pamiętaj, że chcemy uzyskać estymację jak najbliższą prawdziwej wartości parametru, który estymujemy. Taki błąd możemy teoretycznie mierzyć np. jako podniesioną do kwadratu różnicę pomiędzy estymowanym parametrem a naszą estymacją[11]. Nasz błąd wyrazimy więc wzorem

$$MSE = (\hat{\theta} - \theta)^2$$

gdzie $\hat{\theta}$ to nasza estymacja, a θ to prawdziwa wartość estymowanego parametru. Zauważ, że nasz estymator $\hat{\theta}$ jest zmienną losową (zależy od losowej próbki na której go policzymy),

[10] Zachęcam do lektury bardziej zaawansowanych źródeł: np. `https://en.wikipedia.org/wiki/Maximum_likelihood_estimation#Properties`

[11] błąd ten ma wiele użytecznych własności o których będziemy się jeszcze uczyli.

a θ jest stałe (oczywiście nieznamy np. średniej wysokości wszystkich ludzi na świecie, ale jednak wiemy, że jest to jakaś konkretna wartość).

Spróbujmy więc pokazać ile wynosi wartość oczekiwana błędu (jakiego błędu się spodziewamy) dla naszej estymacji. W celu uproszenia dalszego zapisu wprowadźmy oznaczenie: $\mathbb{E}\,\hat{\theta} = \bar{\theta}$. Notabene, $\mathbb{E}\,\hat{\theta}$ też nie jest zmienną losową, a jakąś konkretną liczbą.

$$\begin{aligned}
\mathbb{E}[MSE] &= \mathbb{E}\left[(\hat{\theta}-\theta)^2\right] = \mathbb{E}\left[\left((\hat{\theta}-\bar{\theta})+(\bar{\theta}-\theta)\right)^2\right] \text{ /odejmij i dodaj } \bar{\theta}\text{/} \\
&= \mathbb{E}\left[(\hat{\theta}-\bar{\theta})^2\right] + 2\mathbb{E}\left[(\hat{\theta}-\bar{\theta})(\bar{\theta}-\theta)\right] + \mathbb{E}\left[(\bar{\theta}-\theta)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta}-\bar{\theta})^2\right] + 2(\bar{\theta}-\theta)\,\mathbb{E}\left[\hat{\theta}-\bar{\theta}\right] + (\bar{\theta}-\theta)^2 \text{ /}(\bar{\theta}-\theta) \text{ jest stałą/} \\
&= \mathbb{E}\left[(\hat{\theta}-\bar{\theta})^2\right] + 2(\bar{\theta}-\theta)(\bar{\theta}-\bar{\theta}) + (\bar{\theta}-\theta)^2 \text{ /podstawiamy } \mathbb{E}\,\hat{\theta} = \bar{\theta} \text{ /} \\
&= \mathbb{E}\left[(\hat{\theta}-\bar{\theta})^2\right] + (\bar{\theta}-\theta)^2 \\
&= \mathbb{D}^2[\hat{\theta}] + B(\hat{\theta})^2
\end{aligned}$$

gdzie $\mathbb{D}^2[\hat{\theta}]$ to wariancja estymatora, a $B(\hat{\theta})^2$ to jego obciążenie podniesione do kwadratu (porównaj z Definicją 5.5).

Z wyniku naszych obliczeń wynika, że aby spodziewać się jak najmniejszego błędu należy minimalizować z jednej strony wariancję estymatora, a z drugiej strony jego obciążenie. Obserwujemy także pewnego rodzaju przetarg pomiędzy wariancją a obciążeniem: jeśli jesteśmy w stanie zamienić estymator nieobciążony na inny, o małym obciążeniu, za to z dużo mniejszą wariancją to wyjdziemy na tym na plus!

■ **Przykład 5.5** Załóżmy, że mamy dwa estymatory. Pierwszy z nich jest nieobciążony o wariancji wynoszącej 10, drugi zaś jest estymatorem obciążonym o obciążeniu równym 2 i wariancji wynoszącej 4.

$$\mathbb{E}[MSE(\hat{\theta}_1)] = \mathbb{D}^2[\hat{\theta}_1] + B(\hat{\theta}_1)^2 = 10 + 0 = 10$$

$$\mathbb{E}[MSE(\hat{\theta}_2)] = \mathbb{D}^2[\hat{\theta}_2] + B(\hat{\theta}_2)^2 = 4 + 2^2 = 8$$

Zauważ, że w rozważanej sytuacji spodziewany błąd drugiego estymatora (przypomnijmy: obciążonego!) jest mniejszy niż spodziewany błąd estymatora nieobciążonego. ■

Powyższy przetarg (i nasze równanie) jest tak słynne, że aż doczekało się swojej własnej nazwy *the bias-variance tradeoff* i własnej strony na Wikipedii[12]. Po polsku nomenklatura nie jest jeszcze ustalona (przynajmniej według najlepszej wiedzy autora) .

Literatura

Literatura powtórkowa

Zaawansowany opis różnych sposobów generowania liczb losowych, wraz z analizami ich działami, można znaleźć w rozdziale 4 książki [2]. Opis estymatorów punktowych można znaleźć w podrozdziale 2.2 książki [4]. Analizę estymatorów uzyskanych za pomocą metody Monte Carlo można znaleźć w rozdziale 8 książki [3].

[12] `https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff`

Literatura dla chętnych

Metody Monte Carlo są bardzo często używane w praktyce. Można je stosować np. w planowaniu projektów, gdzie mamy zbiór zadań, które muszą się wykonywać jedno po drugim. Zadania te (jak w życiu) mogą trwać dłużej lub krócej niż się planuje, a ewentualne opóźnienie jednego z nich wpływa na wszystkie następne. Filmik `https://www.youtube.com/watch?v=cRWCwRn8m2k` pokazuje przykładowe zastosowanie metod Monte Carlo do estymacji czasu trwania projektu.

Metody Monte Carlo mają także zastosowanie w robotyce do lokalizacji robotów (ale także w lokalizowaniu piłkarzy na boisku). Filmik `https://www.youtube.com/watch?v=sz7cJuMgKFg` omawia metodę filtru cząsteczkowego (Monte Carlo Localization) do ustalenia pozycji pająka na mapie, mając do dyspozycji robota, który jeździ po mapie wykonując pomiary odległości - pomiary te (jak w życiu) są niedokładne, stąd zadanie jest trudniejsze niż się wydaje. Można zobaczyć także projekt studencki, który wykorzystuje to w praktyce `https://www.youtube.com/watch?v=n60FE7izgUo`.

Pytania sprawdzające zrozumienie

Pytanie 5.1 Czym różnią się od siebie estymatory zgodne, nieobciążone, obciążone, efektywne? Zadania na rozumienie definicji tych estymatorów.

Pytanie 5.2 Czy znasz wzór na wartość oczekiwaną i odchylenie standardowe średniej arytmetycznej $\bar{X}$ z próby? Czy rozumiesz dlaczego wariancja $\mathbb{D}^2[\bar{X}]$ jest mniejsza niż wariancja $\mathbb{D}^2[X]$?

Pytanie 5.3 Czy umiesz pokazać, że estymator wariancji $\frac{1}{n-1}\sum(X_i - \bar{X})^2$ jest nieobciążony?

Bibliografia

[1] Metoda Monte Carlo – Wikipedia, wolna encyklopedia. `https://pl.wikipedia.org/wiki/Metoda_Monte_Carlo`. Dostęp: 2016-02-10.

[2] Siegmnud Brandt. *Analiza danych*. Wydawnictwo Naukowe PWN, 2002.

[3] Jacek Koronacki i Jan Mielniczuk. *Statystyka dla studentów kierunków technicznych i przyrodniczych*. Wydawnictwo Naukowo-Techniczne, 2006.

[4] W. Krysicki, J. Bartos, W. Dyczka, K. Królikowska, i M. Wasilewski. *Rachunek prawdopodobienstwa i statystyka matematyczna w zadaniach. Część II: Statystyka matematyczna*. Wydawnictwo Naukowe PWN, 2002. ISBN 8301113847.