Zadanie 1:
Obserwujemy dwie niezależne próby losowe $(X_1, \ldots, X_n)$, $(Y_1, \ldots, Y_m)$. Wiadomo, że $X_i \sim N(2\mu, 1)$ oraz $Y_i \sim N(\mu, 1)$.

- Wyznaczyć Metodą Największej Wiarogodności estymator parametru μ. Czy otrzymany estymator jest nieobciążony?
- Wyznaczyć ryzyko (średni błąd kwadratowy) uzyskanego estymatora.

Rozwiązanie: *W rozwiązaniu popełniono literówkę: próba dla Y ma n a nie m obserwacji. Schemat rozwiązania jest poprawny, po poprawieniu literówki otrzymuje się oczekiwane odpowiedzi.*

a. Oznaczmy estymator największej wiarygodności $\hat{\mu}$, funkcję wiarygodności L, funkcję log-wiarygodności $l = l(X_1, ..., X_n, Y_1, ..., Y_n, m)$, a gęstość prawdopodobieństwa rozkładu normalnego o średniej μ i odchyleniu standardowym 1 przez p_μ.

Warunek: $\frac{\partial l}{\partial m} = 0$ i $\frac{\partial^2 l}{\partial m^2} > 0$.

$L(...) = \prod_{i=1}^n p_m(X_i) \prod_{i=1}^n p_{2m}(Y_i) = (\frac{1}{\sqrt{2\pi}})^{2n} \prod_{i=1}^n \exp(\frac{-(X_i-2m)^2}{2}) \prod_{i=1}^n \exp(\frac{-(Y_i-m)^2}{2})$

$l(...) = -n\ln(2\pi)\frac{1}{2}\sum_{i=1}^n(-(X_i-2m)^2 - (Y_i-m)^2) = \ln(2\pi)\frac{n}{2}\sum_{i=1}^n((X_i-2m)^2 + (Y_i-m)^2)$

$\frac{\partial l}{\partial m} = C\sum_{i=1}^n[2(X_i-2m)(-2) + 2(Y_i-m)(-1)] = C\sum_{i=1}^n(-4X_i - 2Y_i) + 10nm = 0$ (C – stała)

Stąd: $m = \frac{1}{5n}\sum_{i=1}^n(2X_i + Y_i)$ to kandydat na estymator największej wiarygodności parametru μ.

Oczekiwana odpowiedź bez literówki to $m = \frac{1}{4n+m}(\sum_{i=1}^n 2X_i + \sum_{i=1}^m Y_i)$.

$\frac{\partial^2 l}{\partial m^2} = \frac{\partial}{\partial m}[C\sum_{i=1}^n(-4X_i - 2Y_i) + 10nm] = 10Cn$. $C = \frac{n\ln(2\pi)}{2} > 0$. Zatem m w postaci jak wyżej **jest** estymatorem największej wiarygodności parametru μ.

Obciążenie estymatora: $\mathbb{E}[m] - \mu = \mathbb{E}[\frac{1}{5n}\sum_{i=1}^n(2X_i + Y_i)] - \mu = \frac{1}{5n}\sum_{i=1}^n(2\,\mathbb{E}[X_i] + \mathbb{E}[Y_i]) - \mu = \frac{1}{5n}n(2(2\mu) + \mu) - \mu = \frac{5\mu}{5} - \mu = 0$. Zatem jest to estymator nieobciążony.

b. $MSE := \mathbb{E}[(m-\mu)^2] = \mathbb{E}[m^2] - 2\,\mathbb{E}[m]\mu + \mu^2 = \mathbb{E}[m^2] - \mu^2$.

$\mathbb{E}[m^2] = \frac{1}{25n^2}\mathbb{E}[(\sum_{i=1}^n(2X_i+Y_i))^2] = \frac{1}{25n^2}(\sum_{i\neq j}\mathbb{E}[(2X_i+Y_i)(2X_j+Y_j)] + \sum_{i=1}^n \mathbb{E}[(2X_i+Y_i)^2]) = \frac{1}{25n^2}(\sum_{i\neq j}(4\,\mathbb{E}[X_iX_j] + 2\,\mathbb{E}[X_iY_j] + 2\,\mathbb{E}[Y_iX_j] + \mathbb{E}[Y_iY_j]) + \sum_{i=1}^n(4\,\mathbb{E}[X_i^2] + 4\,\mathbb{E}[X_iY_i] + \mathbb{E}[Y_i^2]))$.

- $\mathbb{E}[X_iX_j] =$ (z niezależności) $= \mathbb{E}[X_i]\,\mathbb{E}[X_j] = 4\mu^2$
- $\mathbb{E}[X_iY_j] = \mathbb{E}[Y_iX_j] = \mathbb{E}[X_iY_i] = 2\mu^2$
- $\mathbb{E}[Y_iY_j] = \mu^2$
- $\mathbb{E}[X_i^2] = \mathbb{E}[(X_i-2\mu)^2 + 4X_i\mu - 4\mu^2] = \mathrm{Var}[X_i] + 4\,\mathbb{E}[X_i]\mu - 4\mu^2 = 1 + 8\mu^2 - 4\mu^2 = 1 + 4\mu^2$
- $\mathbb{E}[Y_i^2] = \mathbb{E}[(Y_i-\mu)^2 + 2Y_i\mu - \mu^2] = \mathrm{Var}[Y_i] + 2\,\mathbb{E}[Y_i]\mu - \mu^2 = 1 + \mu^2$

Stąd $\mathbb{E}[m^2] = \frac{1}{25n^2}(\sum_{i\neq j}(4\times 4\mu^2 + 2\times 2\mu^2 + 2\times 2\mu^2 + \mu^2) + \sum_{i=1}^n(4(1+4\mu^2) + 4\times 2\mu^2 + 1 + \mu^2)) = \frac{1}{25n^2}(\sum_{i\neq j}25\mu^2 + \sum_{i=1}^n(5 + 25\mu^2)) = \frac{1}{25n^2}((n^2-n)25\mu^2 + n(5+25\mu^2)) = \frac{1}{25n}(25n\mu^2 - 25\mu^2 + 5 + 25\mu^2) = \mu^2 + \frac{5}{25n} = \mu^2 - \frac{1}{5n}$.

Stąd $MSE = \frac{1}{5n}$.

Oczekiwana odpowiedź bez literówki to $\frac{1}{4n+m}$.

Zadanie 2:
Niech $(X_1, \dots, X_n)$ będą niezależnymi zmiennymi losowymi o takim samym rozkładzie o gęstości postaci:

$$f_\lambda(x) = \frac{1}{2\lambda^3} x^2 e^{-\frac{x}{\lambda}}, x > 0, \lambda > 0$$

.

- Wyznacz estymator Metodą Największej Wiarogodności nieznanego parametru λ.
- Wiedząc, że wartość oczekiwana wynosi $\mathbb{E}X_i = 3\lambda$, sprawdź, czy otrzymany estymator jest estymatorem nieobciążonym.
- Wyznacz ryzyko dla otrzymanego estymatora. Czy otrzymany estymator jest zgodny?
- Czy otrzymany estymator jest efektywny? *Wskazówka:* Skorzystaj z nierówności Craméra-Rao.

Nierówność Craméra-Rao: Załóżmy, że θ jest nieznanym parametrem, który jest szacowany na podstawie n-elementowej próby x_i z rozkładu prawdopodobieństwa o gęstości $f_\theta(x)$. Wariancja dowolnego nieobciążonego estymatora $\hat{\theta}$ parametru θ jest wtedy ograniczona z dołu przez odwrotność informacji Fishera:

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I_n(\theta)}.$$

Informacja Fishera dla n-elementowej próby:

$$I_n(\theta) = n\mathbb{E}\left[\left[\frac{\partial}{\partial\theta} \ln f_\theta(x)\right]^2\right].$$

Rozwiązanie: 1) Metoda największej wiarygodności parametru λ

$$L(X_1, ..., X_2, \lambda) = (\frac{1}{2\lambda^3})^n \prod (X_i^2) exp(-\frac{1}{\lambda}\sum X_i)$$

$$l(X_1, ..., X_2, \lambda) = l(ln(L)) = -n * ln((2\lambda^3)) + 2\sum ln(X_i) - \frac{1}{\lambda}\sum X_i$$

Więc dla maksymalizowania funkcji l

$$\frac{\partial l}{\partial \lambda} = -n * \frac{3}{\lambda} + \frac{1}{\lambda^2}\sum X_i$$

$$\frac{\partial l}{\partial \lambda} = 0 \to n\frac{3}{\lambda} = \frac{1}{\lambda^2}\sum X_i$$

$$\lambda = \frac{1}{3n}\sum X_i$$

jest to estymator największej wiarygodności parametru lambda i przyjmuje swoje maksimum w właśnie $\lambda = \frac{1}{3n}\sum X_i$

2) Niech $EX_i = 3\lambda$, czy estymator jest obciążony?

$$E(\lambda(X_1, ..., X_n)) = E(\frac{1}{3n}\sum X_i) = \frac{1}{3n}\sum E(X_i) = \frac{1}{3n}\sum 3\lambda = \frac{1}{3n}3n\lambda = \lambda$$

W związku z czym jest nieobciązonym estymatorem.

3) Ryzyko dla estymatora (średni błąd kwadratowy). Czy Estymator zgodny?

Średni błąd kwadratowy to $b(\theta_n)^2 + Var(\theta_n)$, gdzie $b(\theta_n)$ to obciążenie estymatora.

Skoro estymator nie jest obciążony to obciążenie jest równe 0. Wariancja zaś jest równa:

$$Var(\lambda(X_1,...,X_n)) = E(\lambda(X_1,...,X_n)^2) - E(\lambda(X_1,...,X_n))^2 = E(\frac{1}{9n^2}(\sum X_i)^2) - \lambda^2 = \frac{1}{9n^2}E((\sum X_i)^2) - \lambda^2$$

co będzie równe

$$\frac{1}{9n^2}E(\sum\sum X_i X_j) - \lambda^2 = \frac{1}{9n^2}(\sum\sum E(X_i)E(X_j)) - \lambda^2 = \frac{1}{9n^2}(n * n * \mu^2) - \lambda^2$$

co się równa:

$$\mu^2 - \lambda^2$$

gdzie μ jest wartością oczekiwaną zmiennej X_i

Czy estymator jest zgodny? Estymator zgodny jest gdy dla każdego $\epsilon > 0$

$$\lim_{n\to\infty} P(|\lambda_n - \lambda| \geq \epsilon) = 0$$

Inaczej dla $\lambda(X_1,...,X_n)$

$$\lim_{n\to\infty} P(|\frac{1}{3n}\sum X_i - \lambda| \geq \epsilon) = 0$$

z nierówności Czybyszewa

$$P(|\frac{1}{3n}\sum X_i - \lambda| \geq t(\mu^2 - \lambda^2)) \leq \frac{1}{t^2}$$

Obserwacja: dla estymatorów zgodnych $\lim_{n\to\infty} R(\theta) \to 0$ *Nasz estymator jest zgodny.*

4. Estymator jest nieobciążony, więc możemy rozważać efektywność. Na podstawie nierówności Cramera-Rao znajdujemy minimalną wariancję – nasz estymator taką ma, więc jest efektywny.

Zadanie 3:

Niech $X_1, \ldots, X_n$ będą próbą losową z rozkładu normalnego $N(\mu, \sigma^2)$.

- Na podstawie nierówności Craméra-Rao wyznacz dolne ograniczenie dla wariancji nieobciążonego estymatora μ.
- Czy estymator Metody Największej Wiarogodności dla μ jest estymatorem efektywnym? *Wskazówka:* skorzystaj z wyników zadania 3.7.
- Czy średnia arytmetyczna obserwacji o numerach nieparzystych jest efektywnym estymatorem parametru μ? Zakładamy, że liczba obserwacji jest parzysta.

Rozwiązanie:) $\mathrm{Var}(\hat{\mu}) \geq \frac{1}{I_n(\hat{\mu})}$,

$f_{\hat{\mu}}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\hat{\mu})^2}{2\sigma^2})$

$\log f_{\hat{\mu}}(x) = -\log \sigma\sqrt{2\pi} - \frac{(x-\hat{\mu})^2}{2\sigma^2}$

$\frac{\partial}{\partial \hat{\mu}} \log f_{\hat{\mu}}(x) = \frac{x-\hat{\mu}}{\sigma^2}$

$\frac{\partial^2}{\partial \hat{\mu}^2} \log f_{\hat{\mu}}(x) = \frac{-1}{\sigma^2}$

$-\mathbb{E}\frac{\partial^2}{\partial \hat{\mu}^2} \log f_{\hat{\mu}}(x) = -\mathbb{E}\frac{-1}{\sigma^2} = \frac{1}{\sigma^2}$, skąd z nierówności C-R:

$\mathrm{Var}(\hat{\mu}) \geq \frac{\sigma^2}{n}$

b)
Z podpowiedzi, MLE:
$\hat{\mu} = \overline{X_n}$
$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$
Estymator efektywny, to estymator o najmniejszej możliwej wariancji.
Z mojego rozwiązania zadania zad.9 z serii 1 z podpunktu b):

$\mathrm{Var}(\overline{X}) = \frac{1}{n}\sigma^2 + \frac{2}{n^2}\sum_{i \neq j}\mathrm{Cov}(x_i, x_j)$, ponieważ X_i pochodzą z próby, to są niezależne $\Rightarrow Cov(x_i, x_j) = 0$, dla wszystkich i, j.

Czyli $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$, więc na mocy poprzedniego zadania jest estymatorem efektywnym (Średnia z próbki jest estymatorem nieobciążonym, więc możemy zastosować nierówność C-R).

c)
Zauważmy, że możemy przenumerować próbkę tak aby obserwacje o numerach nieparzystych po przenumerowaniu miały numery od 1 do $\frac{n}{2}$, wtedy:
$\mathbb{E}\overline{X}^{np}_{\frac{n}{2}} = \frac{\sum_{i=1}^{\frac{n}{2}}\mathbb{E}X_i^{np}}{\frac{n}{2}} = \mu$
i możemy skorzystać ze wzoru z zadania 9 z poprzedniej serii przyjmując $\overline{X} = \overline{X}^{np}_{\frac{n}{2}}$ i pamiętając o niezależności, otrzymujemy:

$\mathrm{Var}(\overline{X}^{np}_{\frac{n}{2}}) = \frac{2}{n}\sigma^2 > \frac{\sigma^2}{n}$, o ile $\sigma^2 \neq 0$

Czyli, obserwacje o numerach nieparzystych nie są efektywnym estymatorem o ile $\sigma^2 > 0$, w przeciwnym przypadku (dla $\sigma^2 = 0$), jest to efektywny estymator.

Zadanie 4:
Niech $X_1, \ldots, X_n$ będą próbą losową z rozkładu jednostajnego $U(-\theta, 3\theta)$, a $Y_1, \ldots, Y_n$ z rozkładu jednostajnego $U(-2\theta, 6\theta)$, gdzie $\theta > 0$. Obie próby są niezależne.

- Jaki warunek muszą spełniać a i b, aby estymator postaci $\hat{\theta} = a\bar{X} + b\bar{Y}$ był nieobciążonym estymatorem parametru θ? Dla $a = b = 1$ wyznacz ryzyko (błąd średniokwadratowy) tego estymatora.
- Na podstawie próby $Y_1, \ldots, Y_n$ badacz postanowił oszacować θ, stosując następującą procedurę. Drugi moment centralny (wariancję) z próby postanowił porównać z wariancją z rozkładu teoretycznego i na tej podstawie znaleźć wynik. Oszacuj θ, stosując opisaną procedurę.

Rozwiązanie:

a) Liczę wartość oczekiwaną estymatora:

$$E(a\bar{X} + b\bar{Y}) = aE(\bar{X}) + bE(\bar{Y}) = a\frac{1}{n}\sum_{i=1}^{n} E(X_i) + b\frac{1}{n}\sum_{i=1}^{n} E(Y_i) = aE(X_1) + bE(Y_1) =$$

$$\frac{a(-\theta + 3\theta) + b(-2\theta + 6\theta)}{2} = \frac{2a\theta + 4b\theta}{2} = a\theta + 2b\theta$$

Zatem, warunek który muszą spełnić a i b aby estymator był nieobciążony to:

$$a\theta + 2b\theta = \theta \Leftrightarrow a = 1 - 2b$$

Błąd średniokwadratowy: Jeżeli $a = b = 1$ to estymator ma postać $\hat{\theta} = \bar{X} + \bar{Y}$. Zaczynamy od wyznaczenia wartości oczekiwanej: $\mathbb{E}\hat{\theta} = \mathbb{E}X_1 + \mathbb{E}Y_1 = \theta + 2\theta = 3\theta$.

Obciążenie estymatora wynosi: $b(\hat{\theta}) = 2\theta$. Wariancja estymatora: $\operatorname{Var}\hat{\theta} = \operatorname{Var}\bar{X} + \operatorname{Var}\bar{Y} = \frac{1}{n}\operatorname{Var}X_1 + \frac{1}{n}\operatorname{Var}Y_1 = \frac{1}{n}\left(\frac{(3\theta+\theta)^2}{12} + \frac{(6\theta+2\theta)^2}{12}\right) = \frac{20}{3n}\theta^2$.

Ryzyko: $R(\hat{\theta}) = \operatorname{Var} th\hat{e}ta + b(\hat{\theta})^2$

b) Przyrównuję obie wariancje:

$$\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{16\theta^2}{3}$$

Upraszczam prawą stronę równania:

$$\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{\sum_{i=1}^{n} Y_i^2 - 2\bar{Y}\sum_{i=1}^{n} Y_i + n\bar{Y}^2}{n-1} = \frac{\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2}{n-1}$$

Zatem, jako że $\theta > 0$, otrzymuję estymator:

$$\hat{\theta} = \sqrt{\frac{3}{16}\frac{\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2}{n-1}}$$

W rozwiązaniu zabrakło uzasadnienia, czemu wybrano do porównania tę postać estymatora wariancji.

Zadanie 5:
Niech X będzie zmienną losową z rozkładu o gęstości $f_\theta(x) = \theta x^{\theta-1}$ dla $x \in [0, 1]$, gdzie θ jest nieznanym parametrem. Niech c będzie ustaloną dodatnią liczbą. Test jest zbudowany w następujący sposób: jeśli $X \geq c$, to należy przyjąć $H_1 : \theta = 4$, a gdy $X < c$ to przyjmujemy $H_0 : \theta = 2$. Wyznaczyć:

- prawdopodobieństwo popełnienia błędów pierwszego i drugiego rodzaju oraz moc testu;
- wartość c, przy której suma prawdopodobieństw błędów pierwszego i drugiego rodzaju jest najmniejsza.

Rozwiązanie: Na początek prawdopodobieństwo popełnienia błędu pierwszego rodzaju.

Jest to prawdobieństwo, że odrzucimy H_0, pod warunkiem, że jest to hipoteza prawdziwa. To znaczy, że jest to prawdopodobieństwo, że otrzymamy $X \geq c$, kiedy $\theta = 2$.

Rozkład o gęstości $f_2(x)$ na przedziale [0,1] to po prostu rozkład o gęstości 2x na przedziale [0,1].

Prawdopodobieństwo, że X wylosowane z tego rozkładu będzie większe bądź równe c, jest równe zero, gdy c jest większe od 1, ponieważ losujemy na przedziale [0,1].

Dla c należącego do (0,1], to prawdopodobieństwo będzie równe $1 - \int_0^c 2x = 1 - c^2$.

Jako, że c jest dodatnie, to c należące do (0,1] i c>1 to wszystkie opcje, które musimy rozważyć.

Teraz prawdopodobieństwo popełnienia błędu drugiego rodzaju.

Jest to prawdopodobieństwo, że przyjmiemy H_0, podczas, gdy nie jest to hipoteza prawdziwa. To znaczy, że jest to prawdopodobieństwo, że otrzymamy X < c, kiedy $\theta = 4$.

$f_4(x) = 4x^3$. A więc taka jest gęstość naszego rozkładu na przedziale [0,1].

Dla c>1 prawdopodobieństwo będzie, oczywiście, wynosiło 1. Każde X z przedziału [0,1] będzie mniejsze od takiego c, niezależnie, od rozkładu na tym przedziale.

Dla c należącego do (0,1], to prawdopodobieństwo to będzie $\int_0^c 4x^3 = c^4$.

Moc testu to 1-prawdopodobieństwo popełnienia błędu drugiego rodzaju, czyli, u nas, $1 - c^4$ dla c na przedziale (0,1], i 0 dla c większego od 1.

Suma prawdopodobieństwa błędów pierwszego i drugiego rodzaju to 1 dla c>1, i $1 - c^2 + c^4$ dla c należącego do (0,1]. Znajdźmy minimum funkcji $1 - c^2 + c^4$ na przedziale (0,1]. Będzie to albo w 1, albo w dążeniu do 0, albo gdzieś, gdzie pochodna tej funkcji się zeruje. W c=1 ta funkcja ma wartość 1; w dążeniu c do 0 ta funkcja dąży do 1; poszukajmy pochodnych równych 0.

$$(c^4 - c^2 + 1)' = 4c^3 - 2c = 2c(2c^2 - 1)$$

Pochodna jest równa 0 dla c=0, które nie należy do naszego przedziału, albo dla $c = -\sqrt{1/2}$, które nie należy do naszego przedziału, albo dla $c = \sqrt{1/2}$, które do przedziału należy. A więc funkcja $1 - c^2 + c^4$ albo ma minimum równe 1, albo ma jakieś bardziej ekscytujące minimum dla $c = \sqrt{1/2}$.

Dla $c = \sqrt{1/2}$, $1 - c^2 + c^4 = 1 - 1/2 + 1/4 = 3/4$. To nasze bardziej ekscytujące minimum!

A więc, suma prawdopodobieństw pierwszego i drugiego stopnia będzie najmniejsza dla $c = \sqrt{1/2}$.

Zadanie 6:
Niech X_1, X_2, X_3, X_4 będzie próbą losową z rozkładu $N(\mu, \sigma^2)$. Testujemy hipotezę $H_0 : \sigma^2 = 2$ przeciwko $H_1 : \sigma^2 > 2$. Odrzucamy H_0, jeśli wartość statystyki $\sum_{i=1}^4 (X_i - \bar{X})^2$ przekroczy c.

- Dla jakieś wartości c prawdopodobieństwo błędu I rodzaju wynosi 0,1?
- Podać moc testu dla hipotezy alternatywnej $H_1 : \sigma^2 = 10,7$.

Rozwiązanie: Prawdopodobieństwo błędu I rodzaju: $\alpha = P(T > c|H_0) = 0.1$. Zauważmy, że statystyka $chi^2 = \frac{1}{2}T$ ma rozkład χ^2 o 3 stopniach swobody. Dla poziomu istotności α i hipotezy alternatywnej H_1 obszar krytyczny tej statystyki to $W = [q_{\chi^2}(1-\alpha, 3), \infty)$, gdzie $q_{\chi^2}(p, d)$ – kwantyl rzędu p rozkładu χ^2 o d stopniach swobody. Dla $\alpha = 0.1$ otrzymujemy $q_{\chi^2}(1 - \alpha, 3) = 6.251$. Stąd mamy $c = 12.50$.

b. Niech $\sigma_2^2 = 10.7$.

Moc testu: $1 - \beta$, gdzie β – prawdopodobieństwo błędu II rodzaju: $\beta = P(T \leq c|H_1)$. Przy hipotezie $H_1 : \sigma^2 = 10.7$ prawdziwej statystyka $chi_2^2 = \frac{T}{\sigma_2^2}$ ma rozkład χ^2 o 3 stopniach swobody. Zauważmy, że $chi_2^2 = \frac{2}{\sigma_2^2} chi^2$. Obszar krytyczny: $chi^2 > 6.251$ co jest równoważne $chi_2^2 > 1.168$. Stąd $\beta = F(1.168)$, gdzie F – dystrybuanta rozkładu χ^2 o 3 stopniach swobody. $\beta = 0.2393$, co daje moc testu: 0.7607.

Zadanie 7:
Pewnej grupie 12 pacjentów leczonych na nadciśnienie podano lek. Wyniki pomiaru ciśnienia krwi w tej grupie były następujące:

	1	2	3	4	5	6	7	8	9	10	11	12
przed lekiem	220	180	270	290	200	300	250	190	220	230	260	270
po leku	190	170	220	260	220	200	260	150	160	170	210	190

- Zakładając, że rozkład ciśnienia krwi jest normalny, zweryfikować hipotezę o nieskuteczności podanego leku. Przyjmij poziom istotności $\alpha = 0,01$.
- Prawdopodobieństwo tego, że lek zadziała jest równe 0,36. Wyznacz takie k, żeby liczba osób, u których lek zadziała w losowej próbie 400 osób należała do przedziału $(144-k, 144+k)$ z prawdopodobieństwem 0,9.

Rozwiązanie: textbfa. Niech rozkład ciśnienia krwi przed lekiem to $\mathcal{N}(\mu_1, \sigma_1)$, po leku to $\mathcal{N}(\mu_2, \sigma_2)$. Niech próba z pierwszego rozkładu to $(X_1, ..., X_n)$, a z drugiego to $(Y_1, ..., Y_n)$, gdzie $n = 12$.

Hipoteza zerowa: lek jest nieskuteczny, tzn. $\mu_1 \leq \mu_2$. Niech $\mu = \mu_2 - \mu_1$. Wówczas możemy sformułować hipotezę zerową jako: $\mu = 0$ (pomijamy możliwość $\mu > 0$, żeby hipoteza zerowa była dobrze sformułowana).

Hipoteza alternatywna: $\mu < 0$ (ciśnienie spadło).

Dalej wykorzystamy fakt, że rozkład różnic $Y_i - X_i =: Z_i$ to $\mathcal{N}(\mu, \sigma)$, gdzie $\sigma := \sqrt{\frac{\sigma_1^2+\sigma_2^2}{n}}$. σ_1 i σ_2 są nieznane, mamy więc $(Z_1, ..., Z_n)$ – próbę z rozkładu normalnego o nieznanej wariancji.

Dobór statystyki testowej. $T(Z_1, ..., Z_n) = \frac{\bar{Z}}{S/\sqrt{n}}$ ma przy $\mu = 0$ rozkład t-Studenta o $n-1$ stopniach swobody, gdzie $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Z_i - \bar{Z})^2$, a $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$.

Obszar krytyczny: $W \subset \mathbb{R} : P(T \in W|H_0) \leq \alpha$. Stąd $W = (-\infty, q(\alpha)]$, gdzie q – kwantyl rozkładu t-Studenta o $n-1$ stopniach swobody.

Obliczamy wartość statystyki testowej dla zadanej próby:

$Z = (-30, -10, -50, -30, 20, -100, 10, -40, -60, -60, -50, -80)$

$\bar{Z} = -40$. $S = 34.90$. $T = -3.970$. $q(0.01) = -2.718$.

$-3.970 \in (-\infty, -2.718]$, stąd odrzucamy hipotezę zerową. Lek jest skuteczny.

Alternatywne rozwiązanie z wykorzystaniem testu Wilcoksona

Lek jest nieskuteczny, jeśli rozkład ciśnienia przed lekiem jest jednakowy, jak rozkład po leku. Użyję testu Wilcoxona- takie rzeczy to to, do czego służy.

Oto potrzebne nam dane dla naszych par:

i	1	2	3	4	5	6	7	8	9	10	11	12
$\lvert x_{2,i} - x_{1,i}\rvert$	30	10	50	30	20	100	10	40	60	60	50	80
$sgn(x_{2,i} - x_{1,i})$	-1	-1	-1	-1	+1	-1	+1	-1	-1	-1	-1	-1

Jak widać, wszystkie 12 par ma niezerową różnicę. A więc $n_r = 12$. Teraz nadajmy parom rangi, zgodnie z rosnącymi wartościami $|x_{2,i} - x_{1,i}|$, i przypiszmy je wartościom $sgn(x_{2,i} - x_{1,i})$, aby policzyć nasze W.

i	1	2	3	4	5	6	7	8	9	10	11	12
$\lvert x_{2,i} - x_{1,i}\rvert$	30	10	50	30	20	100	10	40	60	60	50	80
$sgn(x_{2,i} - x_{1,i})$	-1	-1	-1	-1	+1	-1	+1	-1	-1	-1	-1	-1
R_i	4.5	1.5	7.5	4.5	3	12	1.5	6	9.5	9.5	7.5	11

$$W = -4.5 - 1.5 - 7.5 - 4.5 + 3 - 12 + 1.5 - 6 - 9.5 - 9.5 - 7.5 - 11 = -69$$

Jako, że n_r jest duże (większe bądź równe 10), to W ma rozkład asymptotycznie normalny. Kwantyl rzędu 0.005- czyli ten, od którego W musi być mniejsze, by być istotne- to -2.58. -69 jest absurdalnie mniejsze od -2.58. Obaliliśmy hipotezę o nieskuteczności podanego leku; lek jest skuteczny, dla naszego poziomu istotności.

b. *W zadaniu można było skorzystać z CTG.*

Rozkład dwumianowy z parametrami $n = 400$ (liczba prób) i $p = 0.36$ (prawdopodobieństwo sukcesu w pojedynczej próbie). Szukamy k takiego, że $F(144+k)-F(144-k) = 0.9$, gdzie F – dystrybuanta tego rozkładu.

$F(x) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$.

$F(144+k) - F(144-k) = \sum_{i=144-k}^{144+k} \binom{n}{i} p^i (1-p)^{n-i}$.

Numeryka daje: $F(144+15) - F(144-15) = 0.881$, $F(144+16) - F(144-16) = 0.904$. Stąd k dające wartość najbliższą poszukiwanej to 16.

Zadanie 8:
Student szuka zajęcia podczas kwarantanny. Z nudów zajął się zbieraniem danych o wszystkim, co tylko przyjdzie mu do głowy. Policzył ziarenka ryżu w 12 torbach, które kupił, robiąc zapasy. Zdziwiło go, że w każdej jest inna liczba ziarenek (mimo że są od jednego producenta!). Wyniki pomiarów zawiera tabela.

1	2	3	4	5	6	7	8	9	10	11	12
8976	8982	8970	8990	8986	9000	8971	8965	8993	8986	8982	8979

- Wyznacz przedział ufności dla średniej liczby ziarenek ryżu na poziomie ufności 0,99. Zakładamy, że liczba ziarenek ryżu pochodzi z rozkładu normalnego $N(\mu, \sigma^2)$, z nieznanymi parametrami μ i σ^2. Zweryfikuj hipotezę, że przeciętna liczba ziarenek ryżu w torebce to 8980.
- Wyznacz przedział ufności dla wariancji liczby ziarenek ryżu w torebce na poziomie ufności 0,99. Zweryfikuj hipotezę, że wariancja liczby ziarenek ryżu wynosi 100.

Rozwiązanie: a) Średnia $= 8981\frac{2}{3}$
Wariancja $= 94\frac{8}{9}$
$\alpha = 0,01$
$n = 12$
Przedział ufności dla średniej:
$(8981\frac{2}{3} - t(1-\frac{0,01}{2}, 11)\frac{\sqrt{94\frac{8}{9}}}{\sqrt{11}}, 8981\frac{2}{3} + t(1-\frac{0,01}{2}, 11)\frac{\sqrt{94\frac{8}{9}}}{\sqrt{11}})$
$1 - \frac{0,01}{2} = 0,995$
Z tablic: $t(0,995, 11) = 3,4966$

Test:
$H_0 : \mu = 8980$
Przeciw: $H_3 : \mu \neq 8980$
Statystyka testowa: $T = \frac{8981\frac{2}{3} - 8980}{\sqrt{94\frac{8}{9}}} \cdot \sqrt{11} = \frac{\frac{5}{3}}{\frac{\sqrt{854}}{3}} \cdot \sqrt{11} = \frac{5}{\sqrt{854}} \cdot \sqrt{11}$
Obszar krytyczny:
$W = (-\infty, -t(1-\frac{\alpha}{2}, 11)) \cup [t(1-\frac{\alpha}{2}, 11), \infty)$
Przyjmijmy: $\alpha = 0,01$
Jak poprzednio z tablic: $t(0,995, 11) = 3,4966$, obszar krytyczny:
$W = (-\infty, -3,4966) \cup [3,4966, \infty)$
Statystyka w przybliżeniu wynosi 0,567 i nie leży w obszarze krytycznym, nie ma podstaw do odrzucenia H_0.

b)
Wariancja.
Przedział ufności ze wzoru dla wariancji:
$(\frac{12\cdot 94\frac{8}{9}}{\chi^2(0{,}995{,}11)}, \frac{12\cdot 94\frac{8}{9}}{\chi^2(0{,}005{,}11)})$

Gdzie z tablic: $\chi^2(0,995,11) = 2,6032$ i $\chi^2(0,005,11) = 26,7569$

Test:
$H_0 : \sigma^2 = 100$
Przeciw: $H_3 : \sigma^2 \neq 100$
Statystyka testowa: $T = \frac{12\cdot 94\frac{8}{9}}{100}$
Obszar krytyczny:
$W = (0, \chi^2(\frac{\alpha}{2}, 11)] \cup [\chi^2(1 - \frac{\alpha}{2}, 11), \infty)$
Przyjmijmy: $\alpha = 0,01$
Jak poprzednio: $\chi^2(0,005,11) = 26,7569$ i $\chi^2(0,995,11) = 2,6032$
Obszar krytyczny: $W = (0; 2,6032] \cup [26,7569, \infty)$
Statystyka w wynosi: $\frac{10248}{900}$, co w przybliżeniu wynosi 11,387 i nie leży w obszarze krytycznym, nie odrzucamy zatem H_0 na rzecz H_3.

Zadanie 9:
Przeprowadzono ankietę, w której dla 21 losowo wybranych studentów średnia liczba filmów obejrzanych w kinie w ciągu roku wyniosła 30, a próbkowe odchylenie standardowe 10, zaś dla losowo wybranych 22 licealistów miary te były równe odpowiednio 25 i 8. Zakładając, że liczba wizyt w kinie ma rozkład normalny, odpowiedz na pytania, weryfikując odpowiednie hipotezy na poziomie istotności $\alpha = 0,05$.

- Czy studenci chodzą do kina tak samo często jak licealiści, czy częściej? A może rzadziej?
- Czy wariancja liczby wizyt w kinie jest taka sama dla licealistów i studentów, czy jest wyższa? A może niższa?

Rozwiązanie: a) Ponieważ zmienne losowe - liczba filmów, obejrzanych przez daną osobę w ciągu roku - mają rozkład normalny, a próbki są nierównoliczne, aby dowiedzieć się czy studenci chodzą do kina średnio tak samo często jak licealiści, czy też nie, zastosujemy test dla dwóch średnich. Dla próbki studentów parametry wynoszą: $n_1 = 21, \mu_1 = 30, \sigma_1 = 10$; dla licealistów: $n_2 = 22, \mu_2 = 25, \sigma_2 = 8$.

Będziemy weryfikować hipotezę $H_0 : \mu_1 = \mu_2$, przeciwko hipotezom alternatywnym: $H_1 : \mu_1 < \mu_2$ oraz $H_2 : \mu_1 > \mu_2$ na poziomie istotności $\alpha = 0,05$.

Różnica średnich z prób ma rozkład

$$N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Jeśli H_0 jest prawdziwa, to nasza statystyka testowa:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Obliczamy ją dla danych z zadania:

$$\frac{5}{\sqrt{\frac{100}{21}+\frac{64}{22}}} = 1,8$$

Dla H_1 obszar krytyczny $W_1 = (-\infty, -q(0,95)] = (-\infty, -1,64]$, a $1,8 \notin W_1$, więc nie ma podstaw do odrzucenia H_0, studenci nie chodzą więc do kina rzadziej niż licealiści.

Dla H_2 obszar krytyczny $W_2 = [q(0,95), \infty) = [1,64, \infty)$, a $1,8 \in W_2$, więc odrzucamy H_0. Studenci chodzą więc do kina częściej niż licealiści.

b) Aby zweryfikować analogiczne hipotezy dla wariancji, użyjemy testu F. Będziemy weryfikować hipotezę $H_0 : \sigma_1^2 = \sigma_2^2$, przeciwko hipotezom alternatywnym: $H_1 : \sigma_1^2 < \sigma_2^2$ oraz $H_2 : \sigma_1^2 > \sigma_2^2$ na poziomie istotności $\alpha = 0,05$.

Naszą statystyką testową jest w tym przypadku statystyka F-Snedecora:

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

która ma rozkład Snedecora z $v_1 = n_1 - 1, v_2 = n_2 - 1$. Obliczamy wartość tej statystyki.

$$F = \frac{100}{64} = 1,56$$

Obszary krytyczne są następujące: Przy $H_1 : W_1 = (0, F_{v_1,v_2}(1-\alpha)]$, przy $H_2 : W_2 = [F_{v_1,v_2}(\alpha), \infty)$. Obliczamy odpowiednie kwantyle, korzystając z tablic:

$$F_{20,21}(0,95) = 2,11$$

$$F_{20,21}(0,05) = 2,096$$

$1,56 \in W_1$, więc w pierwszym przypadku odrzucamy H_0 - wariancja liczby wizyt w kinie dla studentów jest mniejsza niż dla licealistów. $1,56 \notin W_2$, więc nie odrzucamy H_0. Wariancja liczby wizyt w kinie dla studentów nie jest większa niż licealistów.

Zadanie 10:
W obliczu nowego, nieznanego dotąd wirusa, próbowano stworzyć szybszy niż dotychczasowy test wykrywający jego nosicielstwo. Wyniki badań nowym testem uzyskane dla reprezentatywnej próby 1000 osób, wskazały, że wśród osób faktycznie zdrowych nowy test wykazał 680 osób zdrowych i 170 zarażonych; wśród osób faktycznie zarażonych test wskazał 120 osób zarażonych i 30 osób zdrowych.

- Oblicz wartości czułości i swoistości dla nowego testu.
- Jaka jest dokładność oraz *false discovery rate* nowego testu?
- U losowo wybranego obywatela nowy test wykazał obecność wirusa. Jakie jest prawdopodobieństwo, że ta osoba faktycznie jest zarażona?

Rozwiązanie:

X	faktycznie chory	faktycznie zdrowy
model chory	120 = TP	170 = FP
model zdrowy	30 = FN	680 = TN

Dowód. W teście sukcesem będzie wykrycie wirusa. Będziemy używać terminologii

- *prawdziwy pozytywny*: zakażone u których test to potwierdza (120)

- *prawdziwy negatywny*: zdrowe, u których test nie wykrywa wirusa (680),
- *fałszywy pozytywny*: zdrowe u których test wykrywa wirusa (170)
- *fałszywy negatywny*: zakażone u których test nie wykrywa wirusa (30)

.

Ad.(A):

$$\text{czułość} = \frac{\text{liczba prawdziwych pozytywnych}}{\text{liczba prawdziwych pozytywnych} + \text{liczba fałszywych negatywnych}} = \frac{120}{150}$$

$$\text{swoistość} = \frac{\text{liczba prawdziwych negatywnych}}{\text{liczba prawdziwych negatywnych} + \text{liczba fałszywych pozytywnych}} = \frac{680}{850}$$

Ad. (B): *dokładność (accuracy)*

$$ACC = \frac{TP + TN}{N} = \frac{800}{1000}$$

Ad. (C): Oznaczmy przez A zdarzenie polegające na tym, że wybrana osoba jest zarażona, a przez B takie, że wykryto u niej wirusa. Wówczas:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Ponieważ, $\mathbb{P}(A \cap B) = \frac{120}{1000}$, a $\mathbb{P}(B) = \frac{850}{1000} \cdot \frac{170}{850} + \frac{150}{1000} \cdot \frac{120}{150} = \frac{290}{1000}$, to

$$\mathbb{P}(A|B) = \frac{120}{290}$$

□

Zadanie 11:

Student próbuje przewidywać liczbę nowych stwierdzonych przypadków choroby. Codziennie zapisuje w zeszycie swoje przewidywanie (nie zmienia modelu), następnego dnia porównuje je z opublikowanymi danymi. Oto jego wyniki z ostatnich dni:

prognoza	100	120	140	160	180	200	220	240
fakt	98	115	152	150	170	168	249	224

- Oblicz MSE dla okresu prognozy.
- Jakie byłoby MSE dla modelu, w którym prognoza byłaby średnią z dostępnych trzech ostatnich pomiarów? Który z modeli byłby lepszy ze względu na RMSE?
- Zweryfikuj, czy wartości prognozy i wartości rzeczywiste pochodzą z tego samego rozkładu.

Rozwiązanie: Wzór na MSE to:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Mamy więc:

$$MSE = \frac{1}{8}\sum_{i=1}^{8}(y_i - \hat{y}_i)^2 =$$

$$= \frac{1}{8}\left((98-100)^2 + \cdots + (224-240)^2\right) =$$

$$= \frac{1}{8}\left(4 + ... + 256\right) =$$

$$= \frac{1}{8}2494 = 311.75$$

W podpunkcie b) zadania nie dysponujemy danymi, które pozwalają nam obliczyć wartości dla takiego samego horyzontu prognozy jak w a). Dlatego tworzymy model dla 5-ciu obserwacji. Tzn.:

$prognoza_2$	122	139	157	162	195
fakt	150	170	168	249	224

Stad mamy, że:

$$MSE_2 = \frac{1}{5}\sum_{i=1}^{5}(y_i - \hat{y}_i)^2 = 2055.2$$

Żeby porównać oba modele, trzeba też zawęzić horyzont dla pierwszego (obliczamy MSE dla 5 ostatnich obserwacji). Stąd:

$$MSE_1 = \frac{1}{5}\sum_{i=1}^{5}(y_i - \hat{y}_i)^2 = 464.2$$

Aby obliczyć RMSE należy:

$$RMSE = \sqrt{MSE}$$

więc:

$$RMSE_1 = 21.54$$

$$RMSE_2 = 45.33$$

Drugi model popełnia większy błąd. Aby zweryfikować czy wartości prognozy i wartości rzeczywiste są z tego samego rozkładu skorzystamy z testu Pearson's chi-squared test. Przyjmijmy $\alpha = 0.05$. Mamy z niego, że:

$$\chi^2 = \sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{y_i}$$

Obliczamy więc:

$$\chi^2 = \sum_{i=1}^{8}\frac{(\hat{y}_i - y_i)^2}{y_i} =$$

$$= \sum_{i=1}^{8}\frac{(100-98)^2}{98} + ... + \frac{(240-224)^2}{224} =$$

$$= 13.0768$$

Dla $\alpha = 0.05$ oraz $df = 7$ wartość rozkładu chi-kwadrat to: 14.06714. Nie odrzucamy więc hipotezy zerowej, że wartości prognozy i wartości rzeczywste pochodzą z tego samego rozkładu.

Zadanie 12:
Dane są trzy próby:

(a) 6,6 ; 0,6 ; 3,4 ; 1,8 ; -0,6; 4,2 ; 2,6 ; 5,4

(b) 2,62; 0,92; 1,75; 4,84; 0,26; 1,27

(c) 1,5; 0,3; 0,8; 2,0; 2,0; -0,8; 0,6; 0,6; 1,5; 0,3

Wykorzystując test Kołmogorowa zweryfikuj hipotezy:

- Próbę (a) pobrano z populacji o rozkładzie $N(3, 2^2)$, poziom istotności $\alpha = 0,05$.
- Próbę (b) pobrano z populacji o rozkładzie $f(x) = \frac{1}{2}e^{-\frac{x}{2}}, x > 0$, poziom istotności $\alpha = 0,1$.
- Próbę (c) pobrano z populacji o rozkładzie $N(0, 1)$, poziom istotności $\alpha = 0,1$.

Rozwiązanie: Dystrybuanta empiryczna F_n, którą będziemy przybliżać tą właściwą, dana jest wzorem:

$$F_n(x) = \frac{1}{n}\sum_{i=0}^{n} \mathbf{1}_{(-\infty,x)}(X_i) \tag{1}$$

Jest to funkcja skokowa, rosnąca i stała przedziałami. Będziemy nią przybliżać inne funkcje ciągłe i rosnące. Widać więc że największa różnica F_n i F_0 musi znajdować się na krawędzi przedziału stałości F_n. (Gdyby znajdowała się w środku przedziału, to różnica pod jego koniec musi być nie mniejsza.)

Podpunkt(a): Dystrybuanta rozkładu $N(3, 2^2)$ może zostać wyrażona przez funkcję specjalną erf w następujący sposób:

$$F_0(x) = \frac{1}{2}(1 + erf(\frac{x-3}{2\sqrt{2}})) \tag{2}$$

Posortujmy wartości z puli (a) rosnąco oraz zapiszmy wartości F_0, i F_n w tych punktach:

X:	-0,6	0,6	1,8	2,6	3,4	4,2	5,4	6,6
$F_0(X)$:	0,0359	0,1151	0,2743	0,4207	0,5793	0,7257	0,8849	0,9641
$F_n(X)$:	0,1250	0,2500	0,3750	0,5000	0,6250	0,7500	0,8750	1,0000

Rozpiszmy, też różnice F_0 i F_n:

X:	-0,6	0,6	1,8	2,6	3,4	4,2	5,4	6,6
$lim_{x->X^-}\lvert F_0(x) - F_n(x)\rvert$:	0,0359	0,0099	0,0243	0,0457	0,0793	0,1007	0,1349	0,0891
$lim_{x->X^+}\lvert F_0(x) - F_n(x)\rvert$:	0,0890	0,1349	0,1007	0,0793	0,0457	0,0243	0,0099	0,0359

Otrzymujemy więc watorść statystyki $\sqrt{8}D_8 = 0,38155$. Wartość kwantyla rozkładu Kołmogorowa rzędu $D_8(0,95)$ jest równa: 0,454. Co nie daje nam podstaw do odrzucenia hipotezy zerowej.

Podpunkt(b): Dystrybuanta rozkładu $f(x) = \frac{1}{2}e^{-\frac{x}{2}}, x > 0$ może zostać wyrażona w następujący sposób:

$$F_0(x) = \int_0^x \frac{1}{2}e^{-\frac{s}{2}}ds = 1 - e^{-\frac{x}{2}} \tag{3}$$

Posortujmy wartości z puli (b) rosnąco oraz zapiszmy wartości F_0, i F_n w tych punktach:

X:	0,26	0,92	1,27	1,75	2,62	4,84
$F_0(X)$:	0,1219	0,3687	0,4701	0,5831	0,7302	0,9111
$F_n(X)$:	0,1667	0,3333	0,5000	0,6667	0,8333	1,0000

Rozpiszmy, też różnice F_0 i F_n:

X:	0,26	0,92	1,27	1,75	2,62	4,84
$lim_{x->X^-}\lvert F_0(x) - F_n(x)\rvert$:	0,1219	0,2021	0,1367	0,0831	0,0635	0,0777
$lim_{x->X^+}\lvert F_0(x) - F_n(x)\rvert$:	0,0448	0,0354	0,0299	0,0835	0,1032	0,0889

Otrzymujemy więc watorść statystyki $\sqrt{6}D_6 = 0,49504$. Wartość kwantyla rozkładu Kołmogorowa rzędu $D_6(0,90)$ jest równa około: 0,468. Co daje nam podstawy do odrzucenia hipotezy zerowej.

Podpunkt(c): Dystrybuanta rozkładu $N(0,1)$ może zostać wyrażona przez funkcję specjalną erf w następujący sposób:

$$F_0(x) = \frac{1}{2}(1 + erf(\frac{x}{\sqrt{2}})) \tag{4}$$

Posortujmy wartości z puli (c) rosnąco oraz zapiszmy wartości F_0, i F_n w tych punktach:

X:	-2,0	-1,5	-0,8	-0,6	-0,3	0,3	0,6	0,8	1,5	2,0
$F_0(X)$:	0,0228	0,0668	0,2119	0,2743	0,3821	0,6179	0,7257	0,7881	0,9332	0,9773
$F_n(X)$:	0,1000	0,2000	0,3000	0,4000	0,5000	0,6000	0,7000	0,8000	0,9000	1,0000

Rozpiszmy, też różnice F_0 i F_n:

X	-2,0	-1,5	-0,8	-0,6	-0,3	0,3	0,6	0,8	1,5	2,0
$\lvert F_0(X) - F_n(X^-)\rvert$	0,0228	0,0332	0,0119	0,0257	0,0179	0,1179	0,1257	0,0881	0,1332	0,0773
$\lvert F_0(X) - F_n(X^+)\rvert$	0,0772	0,1332	0,0881	0,1257	0,1179	0,0179	0,0257	0,0119	0,0332	0,0228

Otrzymujemy więc wartość statystyki $\sqrt{10}D_{10} = 0,42122$. Wartość kwantyla rozkładu Kołmogorowa rzędu $D_{10}(0,90)$ jest równa około: $0,369$. Co daje nam podstawy do odrzucenia hipotezy zerowej.

Zadanie 13:
Badano związek pomiędzy wykształceniem a zarobkami netto. Każda osoba była pytana o liczbę lat nauki, na jej podstawie sklasyfikowano wykształcenie jako podstawowe (lub mniej), średnie lub wyższe. Zarobki sklasyfikowano na trzech poziomach. Wyniki przedstawia poniższa tabela.

	podstawowe lub mniej	średnie	wyższe
< 2000	54	78	128
2000-4000	75	122	73
> 4000	71	40	49

Dostarczono również fragment dokładnych danych z przeprowadzonej ankiety:

nr res.	1	2	3	4	5	6	7	8	9	10
zarobki	1700	4000	2560	3590	1900	2200	5190	4500	2460	2190
lata nauki	14	16	9	12	8	12	20	13	11	10

- Zweryfikuj hipotezę o niezależności obu cech na poziomie istotności $\alpha = 0,025$.
- Oblicz wartość współczynnika korelacji Spearmana. Czy pomiędzy poziomem wykształcenia oraz zarobkami występuje istotny statystycznie związek? Przyjmij $\alpha = 0.05$.

Rozwiązanie:

a) Żeby ustalić niezależność zmiennych liczę statystykę testu Pearsona:

$$T = \frac{(54 - \frac{260*200}{690})^2}{\frac{260*200}{690}} + \frac{(78 - \frac{260*240}{690})^2}{\frac{260*240}{690}} + \frac{(128 - \frac{260*250}{690})^2}{\frac{260*250}{690}} +$$

$$\frac{(75 - \frac{270*200}{690})^2}{\frac{270*200}{690}} + \frac{(122 - \frac{270*240}{690})^2}{\frac{270*240}{690}} + \frac{(73 - \frac{270*250}{690})^2}{\frac{270*250}{690}} +$$

$$\frac{(71 - \frac{160*200}{690})^2}{\frac{160*200}{690}} + \frac{(40 - \frac{160*240}{690})^2}{\frac{160*240}{690}} + \frac{(49 - \frac{160*250}{690})^2}{\frac{160*250}{690}} = 53.591$$

Statystyka ma rozkład χ^2 z 4 stopniami swobody. Wartość krytyczna dla wybranego poziomu istotności wynosi 11.142. Zatem, należy odrzucić hipotezę zerową o niezależności cech.

b) Liczę współczynnik korelacji rang Spearmana:

$$\sum_{n=1}^{n} d_k = (10-3)^2+(3-2)^2+(5-9)^2+(4-5.5)^2+(9-10)^2+(7-5.5)^2+(1-1)^2+(2-4)^2+(6-7)^2+(8-8)^2 = 80,5$$

Zatem:

$$\rho = 1 - \frac{6 \cdot 80,5}{1000 - 10} = 0.512$$

Statystyka testu t przyjmuje wartość:

$$t = \frac{0.512\sqrt{8}}{\sqrt{1 - (0.512)^2}} = 1,686$$

Jako, że $t(0.975, n-2) = 2,306$ to wartość statystyki nie należy do prawostronnego obszaru odrzucenia. Zatem, należy uznać że nie istnieje istotny statystycznie związek między cechami.

Zadanie 14:

Na podstawie wieloletnich obserwacji, stwierdzono, że waga (w kg) pracowników biura ma w przybliżeniu rozkład normalny $N(75, 4^2)$. Tuż przed przerwą obiadową, winda przewozi jednorazowo 6 osób.

- Określić rozkład średniej wagi osób przewożonych windą przed przerwą obiadową, przy założeniu, że zawsze przewozi po 6 osób. Jakie jest prawdopodobieństwo, że średnia waga tych osób będzie większa niż 73kg?
- Jakie jest prawdopodobieństwo, że zsumowana waga tych osób przekroczy dopuszczalne obciążenie, tzn. będzie większa niż 480kg?

Rozwiązanie:
Niech $\overline{X}$ będzie zmienną losową oznaczającą średnią wage 6 osób. Wiem że, skoro waga z każdej z osób pochodz z rozkładu gaussowskiego $N(75, 4^2)$, to $\overline{X} \approx N(75, \frac{8}{3})$. Czyli to co chemy policzyć to:
$P(\overline{X} > 73) = P(\overline{X} - 75 > -2) = P(\frac{\overline{X}-75}{\sqrt{\frac{8}{3}}} > \frac{-\sqrt{3}}{\sqrt{2}}) = \Phi(\frac{\sqrt{3}}{\sqrt{2}}) \approx 0,89$
W przedostatnim przejściu użyłem symetrii rozkładu normalnego. Ostatecznie szukane prawdopodobieństwo jest równe $0,89$.

Niech Y będzie zmienną losową określającą zsumowaną wage 6 osób. Wówczas $Y \approx N(450, 96)$.
$P(Y > 480) = P(y - 450 > 30) = P(\frac{Y-450}{\sqrt{96}} > \frac{30}{\sqrt{96}}) = 1 - \Phi(\frac{30}{\sqrt{96}}) \approx 1 - \Phi(3,06) \approx 1 - 0,9988 \approx 0,001$. Czyli szukane prawdopodobieństwo wynosi nieco ponad $0,001$

Zadanie 15:
Władze województwa zamierzają sprawdzić skuteczność programu przeciwdziałania przemocy w rodzinie. Na terenie powiatów umieszczone zostały plakaty informujące o możliwości telefonicznego wsparcia psychologicznego. Zmierzono liczbę wykonanych telefonów w poszczególnych powiatach przed wprowadzeniem plakatów oraz po. Wyniki przedstawia tabela.

powiat	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
przed	118	134	130	124	105	130	130	132	123	128	126	140	135	126	132
po	125	132	138	120	125	127	136	139	131	132	135	136	128	127	130

- Korzystając z testu Wilcoxona, sprawdź, czy liczba wykonanych telefonów zmieniła się.
- Oblicz współczynnik korelacji Spearmana. Czy istnieje statystycznie istotny związek pomiędzy liczbą wykonanych telefonów przed programem a po programie?
- Przeprowadzono badanie ankietowe w powiatach, sprawdzające zasięg oddziaływania programu. W każdym z powiatów wylosowano niezależnie 200 osób, którym zadano pytanie, czy widziały plakaty. W powiecie 7 twierdząco odpowiedziało 49 osób. Zbudować przybliżony przedział ufności dla odsetka osób poddanych programowi w powiecie 7 na poziomie ufności 0.9.

Rozwiązanie: Korzystając z metody obliczania testu Wilcoxona zawartego w wykładzie: 1) Oblicz $|x_{2,i} - x_{1,i}|$ dla $1 < i < n$ gdzie x_2 to po a x_1 to przed.

powiat	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
przed	118	134	130	124	105	130	130	132	123	128	126	140	135	126	132
po	125	132	138	120	125	127	136	139	131	132	135	136	128	127	130
różnica	7	-2	8	-4	20	-3	6	7	8	4	9	-4	-7	1	-2
moduł	7	2	8	4	20	3	6	7	8	4	9	4	7	1	2

2) Ponieważ nie ma par które równają się zero to porządkujemy pary po ich modułach rosnąco i nadajemy im rangi

powiat	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
przed	118	134	130	124	105	130	130	132	123	128	126	140	135	126	132
po	125	132	138	120	125	127	136	139	131	132	135	136	128	127	130
różnica	7	-2	8	-4	20	-3	6	7	8	4	9	-4	-7	1	-2
moduł	7	2	8	4	20	3	6	7	8	4	9	4	7	1	2
rangi	10	2,5	11,5	6	15	4	8	10	11,5	6	14	6	10	1	2,5

3) Następnie obliczamy statystykę

$$W = \sum(sgn(x_{2,i} - x_{1,i}) * R_i)$$

gdzie R_i to ranga. Wychodzi nam że $W = 56$

4) Obliczamy

$$Var(W) = \frac{n(n+1)(2n+1)}{6}$$

które wychodzi 1240 oraz jej pierwiastek to w zaokrągleniu 35,214.

5) Ponieważ n dostatecznie duże to możemy obliczyć zmienną losową z

$$z = \frac{W}{\sigma_W}, \sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{6}}$$

Wtedy z ma rozkład $N(0,1)$

$$z = \frac{56}{35,214} = 1,59$$

W związku z czym widać że liczba telefonów zmieniła się i wzrosła.

Punkt 2: Współczynnik korelacji Spearmana.

Uszeregujmy w rangi wartości przed i po

powiat	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
przed	118	134	130	124	105	130	130	132	123	128	126	140	135	126	132
po	125	132	138	120	125	127	136	139	131	132	135	136	128	127	130
ranga przed	2	13	9	4	1	9	9	11,5	3	7	5,5	15	14	5,5	11,5
ranga po	2,5	9,5	14	1	2,5	4,5	12,5	15	8	9,5	11	12,5	6	4,5	7
moduł różnicy	0,5	3,5	5	3	1,5	4,5	3,5	3,5	5	2,5	5,5	2,5	8	1	4,5
kwadrat	0,25	12.25	25	9	2,25	20,25	12,25	12,25	25	6,25	30,25	4,25	64	1	20,25

Następnie współczynnik rand Spearmana wyraża się przez:

$$\rho = 1 - \frac{6\sum d_k^2}{n^3 - n}$$

$\sum d_k^2$ to 244,5

Więc

$$\rho = 1 - \frac{6 * 244,5}{15^3 - 15} = 1 - \frac{1467}{3360} = 1 - 0,4366 = 0,5633$$

Zakładając H_0 prawdziwe, można korzystać ze statystyki:

$$t = \frac{\rho\sqrt{n-2}}{1-\rho^2}$$

Podstawiając tam liczby:

$$t = \frac{0,5633\sqrt{15-2}}{1-0,5633^2} = \frac{0,5633 * 3,606}{0,6827} = 2,97496$$

Co znaczy że istnieje zależność statystyczna pomiędzy danymi przed i po (wartość krytyczna: 2,16).

Podpunkt 3)

Każda osoba ma prawdopodobieństwo zobaczenia plakatu p. Zmienna losowa przyjmuje wartość 1 gdy dana osoba zobaczyła plakat, a 0 gdy nie. W związku z tym $\bar{X} = \frac{m}{n}$ gdzie n to próba, a m to ilość osób które zobaczyły plakat. Poziom ufności to $0,9 = 1 - 0,1$

W związku z tym przedział ufności wygląda w następujący sposób:

$$(\frac{m}{n} - q(00,5)\sqrt{\frac{\frac{m}{n}(1-\frac{m}{n})}{n}} < p < \frac{m}{n} + q(00,5)\sqrt{\frac{\frac{m}{n}(1-\frac{m}{n})}{n}})$$

gdzie q(0,05) to kwantyl rozkładu normalnego na poziomie 0,05.

Po podstawieniu liczb:

$$(\frac{49}{200} - 1,64\sqrt{\frac{\frac{49}{200}(1-\frac{49}{200})}{200}} < p < \frac{49}{200} + 1,64\sqrt{\frac{\frac{49}{200}(1-\frac{49}{200})}{200}})$$

Co wychodzi:

$$(0,19512, 0,29487)$$

I taki jest przedział ufności odsetka osób które zobaczyły plakat.

Zadanie 16:

Mamy sześciościenną kostkę do gry, przy czym nie znamy prawdopodobieństwa wypadnięcia 6, oznaczonego przez p. W celu oszacowania p rzucamy kostką dopóki nie wypadnie 6 i przez Y oznaczamy liczbę wykonanych rzutów. Jednak jeśli w pierwszych k rzutach nie wypadła 6 to przerywamy eksperyment i $Y = k + 1$. Na podstawie n niezależych powtórzeń powyższego eksperymenty wyznacz estymator największej wiarygodności parametru p.

Rozwiązanie:
W dalszej części zadania będe używał, że $p \in (0, 1)$, tzn kiedy pisze nierówność na p mam na myśli w dziedzinie określoności.
Z treści zadania wiemy że Y ma następujący rozkład:
$P(Y = s) = p(1-p)^{s-1}$ dla $s \in \{1, ..., k\}$.
$P(Y = k + 1) = 1 - \sum_{i=1}^{k} p(1-p)^i = (1-p)^k$.
Stad nasza funkcja wiarygodności ma postać:
$L(Y_1, ..., Y_n, p) = P(Y_1 = y_1, ..., Y_n = y_n) =$ (z niezależności) $= P(Y_1 = y_1) * ... * P(Y_n = y_n) = A$.
Nasze zmienne losowe mają wyszczególnione prawdopodobieństwo dla $k + 1$, stąd załóżmy, że ten wynik uzyskaliśmy w (BSO $0 \leq w \leq n$) "w" ostatnich próbach. Wiedząc to podstawiamy i otrzymujemy:
$A = p(1-p)^{y_1-1} * ... * p(1-p)^{y_{n-w}-1} * ((1-p)^k)^w$.
Niech $G(p) = Ln(L(Y_1, ..., Y_n, p))$. Oczywiście, ponieważ logarytm jest funkcja ściśle rosnącą to $G(p)$ przyjmuje maksimum w $p_0 \iff$ funkcja $L(Y_1, ..., Y_n, p)$ przyjmuje maksimum w p_0.
Podstawiając do definicji funkcji G dane otrzymujemy:
$G(P) = (n-w)ln(p) + (\sum_{i=1}^{n-w}(y_i - 1))ln(1-p) + (kw)ln(1-p)$.

Teraz liczymy pochodną funkcji G.
$\frac{\partial G}{\partial p} = \frac{n-w}{p} - \frac{(\sum_{i=1}^{n-w}(y_i-1))+kw}{1-p}$
Chcemy policzyć maksimum stąd z lematu Fermata pochodna, (o ile funkcja jest różniczkowalna jak w naszym przypadku), zeruje się w punkcie przyjmowania maksimum. Po przyrównaniu do 0 dostajemy:
$\frac{\partial G}{\partial p} = 0 \iff p = \frac{n-w}{kw+\sum_{i=1}^{n-w}(y_i)}$.
Pozostaje sprawdzić czy jest to maksimum. Zauważmy, że:
$\frac{\partial G}{\partial p} > 0 \iff p < \frac{n-w}{kw+\sum_{i=1}^{n-w}(y_i)}$
$\frac{\partial G}{\partial p} < 0 \iff p > \frac{n-w}{kw+\sum_{i=1}^{n-w}(y_i)}$
Czyli pochodna zmienia znak z czego wnioskujemy że punkt $p = \frac{n-w}{kw+\sum_{i=1}^{n-w}(y_i)}$ jest maksimum. Stąd nasze wyliczone p jest dokładnie szukanym parametrem największej wiarygodnosci.

Zadanie 17:
Niech $X_1, ..., X_n$ będzie próbą prostą z rozkładu Poissona o intensywności θ
$P(X_i = x) = \frac{\theta^x}{x!}e^{-\theta}$

- Znajdź $\hat{\theta}$ estymator największej wiarogodności parametru θ.
- Oblicz obciążenie oraz wariancję estymatora $\hat{\theta}$, uzyskanego w poprzednim podpunkcie.
- Jak duże powinno być n, żeby błąd średniokwadratowy dla $\theta = 1$ był mniejszy niż $0,01$, gdzie $MSE(\theta) = E_\theta[(\theta - \hat{\theta})^2]$.

Rozwiązanie: Aby policzyć estymator największej wiarygodności potrzebujemy znaleźć miejsce zerowe funkcji $L(X_1, ..., X_n, \theta)$:

$$L(X_1, ..., X_n, \theta) = \prod_{i=1}^{n} \frac{\theta_i^X}{X_i!}e^{-\theta} = e^{-n\theta}\frac{\theta^{\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i!} \tag{5}$$

Rozważmy jej logarytm $l(X_1, ..., X_n, \theta) = log(L(X_1, ..., X_n, \theta))$:

$$l(X_1, ..., X_n, \theta) = log(\theta)\sum_{i=1}^{n} X_i - n\theta - \sum_{i=1}^{n} X_i! \tag{6}$$

Oraz pierwszą i drugą pochodną l:

$$\frac{\partial l(X_1, ..., X_n, \theta)}{\partial \theta} = \frac{1}{\theta}\sum_{i=1}^{n} X_i - n \tag{7}$$

$$\frac{\partial^2 l(X_1, ..., X_n, \theta)}{\partial \theta^2} = -\frac{1}{\theta^2}\sum_{i=1}^{n} X_i \tag{8}$$

Przyrównanie pierwszej pochodnej do zera daje nam $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$. Wartość drugiej pochodnej w tym punkcie nam wartość: $-\frac{1}{\hat{\theta}^2}n\hat{\theta}$. Jest ona ujemna, więc $\hat{\theta}$ jest lokalnym maksimum.

Policzmy wartość oczekiwaną zmiennej losowej $\hat{\theta}$:

$$E[\hat{\theta}] = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = E[X] = \sum_{x=0}^{\infty} x\frac{\theta^x}{x!}e^{-\theta} = \theta \sum_{x'=0}^{\infty} \frac{\theta^{x'}}{x'!}e^{-\theta} = \theta \tag{9}$$

Policzymy najpierw wartość oczekiwaną X^2:

$$E[X^2] = \sum_{x=0}^{\infty} x^2\frac{\theta^x}{x!}e^{-\theta} = \theta \sum_{x'=0}^{\infty}(x'+1)\frac{\theta^{x'}}{x'!}e^{-\theta} = \theta \sum_{x'=0}^{\infty} x'\frac{\theta^{x'}}{x'!}e^{-\theta}+\theta = \theta^2 \sum_{x''=0}^{\infty} \frac{\theta^{x''}}{x''!}e^{-\theta}+\theta = \theta^2+\theta \tag{10}$$

Przyjżyjmy się wartości oczekiwanej $\hat{\theta}^2$:

$$E[\hat{\theta}^2] = E[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_iX_j] = \frac{1}{n^2}E[\sum_{i=1}^{n} X_i^2+\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} X_iX_j] = \frac{1}{n^2}(\sum_{i=1}^{n} E[X_i^2]+\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} E[X_i]E[X_j]) \tag{11}$$

$$E[\hat{\theta}^2] = \frac{1}{n^2}(nE[X^2] + n(n-1)E[X]^2) = \frac{1}{n}(\theta^2 + \theta + (n-1)\theta^2) = \theta^2 + \frac{\theta}{n} \tag{12}$$

Podstawiając do wzoru na wariancję:

$$Var[\hat{\theta}] = E[\hat{\theta}^2] - E[\hat{\theta}]^2 = \theta^2 + \frac{\theta}{n} - \theta^2 = \frac{\theta}{n} \tag{13}$$

Przyjrzyjmy się wzorowi na błąd średniokwadratowy:

$$MSE(\theta) = E_\theta[(\theta - \hat{\theta})^2] = Var[\hat{\theta}] + (b(\hat{\theta}))^2 \tag{14}$$

Gdzie $b(\hat{\theta})$ jest obciążeniem. W naszym przypadku dostajemy więc $MSE(\theta) = \frac{\theta}{n}$. Chcielibyśmy otrzymać $\frac{\theta}{n} < \frac{1}{100}$, co jest równoważne: $100\theta < n$. Przy $\theta = 1$, otrzymujemy więc: $100 < n$.

Zadanie 18:

Niech $X_1, ..., X_n$ będzie próbą prostą z rozkładu Pareto o parametrach $a > 0, \theta > 0$ o gęstości $f_{\theta,a} = \frac{\theta a^\theta}{x^{\theta+1}}1(x > a)$

Znajdź $\hat{\theta}$ oraz $\hat{a}$ estymatory największej wiarogodności parametrów θ oraz a.

Rozwiązanie:

Konstruujemy funkcję wiarygodności i po zlogarytmowaniu jej szukamy jej ekstremum. Będzie ona niezerowa tylko dla wszystkich $X_i > a$, a największą wartość będzie przyjmowała dla największego a (jest ono w liczniku funkcji gęstości), więc estymowane:

$$\hat{a} = \min_{i=1,...,n}(X_i)$$

.

$$L(X_1, ...X_n, \hat{a}, \theta) = \prod_{i=1}^{n} \frac{\theta a^\theta}{X_i^{\theta+1}}$$

$$l = \ln(L) = n \ln \theta + n\theta \ln \hat{a} - (\theta + 1) \sum_{i=1}^{n} \ln X_i$$

$$\frac{\partial l}{\partial \theta} = \frac{n}{\theta} + n \ln \hat{a} - \sum_{i=1}^{n} \ln X_i = 0$$

Wyliczamy stąd:

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \ln X_i - n \ln(\min_{i=1,...,n}(X_i))}$$

Sprawdzamy jeszcze warunek drugiego rzędu:

$$\frac{\partial^2 l}{\partial \theta^2} = -\frac{n}{\theta^2} < 0$$

Widać więc, że dla wyliczonych estymatorów osiągane jest maksimum. Są to estymatory największej wiarygodności.

Zadanie 19:
Rozważmy model Hardy'ego-Weinberga, trzy genotypy AA,aa oraz Aa występują w proporcjach $\theta^2, (1\theta)^2$ oraz $2\theta(1\theta)$ odpowiednio. Na podstawie obserwacji z populacji n elementowej wyznacz estymator największej wiarogodności parametru θ, a następnie oblicz jego obciążenie i wariancję.
Rozwiązanie: *Brak przesłanego zadania od studentów, ale przykład do znalezienia w Niemiro: Rachunek prawdopodobieństwa i statystyka matematyczna; przykład 3.1.2, 3.4.8, 5.3.2*

Zadanie 20:
Niech $X_1, ..., X_n$ będzie próba prostą z rozkładu Log-normalnego o parametrach $\mu, \sigma^2 > 0$, o gęstości

$$f_{\mu,\sigma^2} = \frac{1}{x\sqrt{2\pi}\sigma} exp(-\frac{(ln(x) - \mu)^2}{2\sigma^2})$$

(A) Znajdź $\hat{\mu}, \hat{\sigma}^2$ estymatory największej wiarygodności parametrów μ, σ^2,

(B) Oblicz obciążenie oraz wariancje estymatora $\hat{\mu}$ uzyskanego w poprzednim podpunkcie,

(C) Jak duże powinno być n, żeby błąd średniokwadratowy dla $\mu = 0$, $MSE(0)$, był mniejszy niż 0.01, gdzie $MSE(\mu) = \mathbb{E}_\theta(\mu - \hat{\mu})^2$.

Rozwiązanie:

Dowód. Ad.(A):

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{x_i} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(ln(x) - \mu)^2}{2\sigma^2})$$

$$l(\mu, \sigma^2) = ln(L(\mu, \sigma^2)) = \sum_{i=1}^{n} (-\frac{(ln(x_i) - \mu)^2}{2\sigma^2} - ln(x_i) - ln(\sqrt{2\pi}\sigma))$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} ln^2(x_i) + \frac{\mu}{2\sigma^2} \sum_{i=1}^{n} ln(x_i) + n\frac{\mu^2}{2\sigma^2} - \sum_{i=1}^{n} ln(x_i) - nln(\sqrt{2\pi}\sigma)$$

Obliczymy pochodne, by znaleźć ekstrema:

$$\frac{\partial}{\partial\mu}l(\mu,\sigma) = \frac{1}{\sigma^2}\sum_{i=1}^{n} ln(x_i) - \frac{\mu}{\sigma^2}n = 0$$

Skąd $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} ln(x_i)$

$$\frac{\partial}{\partial\sigma}l(\mu,\sigma^2) = \frac{1}{\sigma^3}\sum_{i=1}^{n} ln^2(x_i) - \frac{2\mu}{\sigma^3}\sum_{i=1}^{n} ln(x_i) + \frac{\mu^2}{\sigma^3}n - \frac{n}{\sigma} = 0$$

skąd wyliczamy:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} ln^2(x_i) - 2\mu\sum_{i=1}^{n} ln^(x_i) + n\mu^2}{n} = \frac{\sum_{i=1}^{n} ln^2(x_i) - \frac{2}{n}(\sum_{i=1}^{n} ln(x_i))^2 + \frac{1}{n}(\sum_{i=1}^{n} ln(x_i))^2}{n}$$

$$= \frac{\left(\sum_{i=1}^{n} ln(x_i) - \frac{1}{n}(\sum_{i=1}^{n} ln(x_i))\right)^2}{n}$$

Formalnie powinniśmy sprawdzić, czy znalezione rozwiązanie to faktycznie maksimum.

Ad. (B):

$$b(\hat{\mu}) = \mathbb{E}(\mu - \hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}ln(x_i) - \mu = \frac{1}{n}n\mu - \mu = 0$$

$$var(\hat{\mu}) = var(\frac{1}{n}\sum_{i=1}^{n} ln(x_i)) = \frac{1}{n^2}\sum_{i=1}^{n} var(ln(x_i)) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

Ad. (C): $\mu = 0$ oraz MSE(0) < 0.01

$$MSE(\mu) = \mathbb{E}(\mu - \hat{\mu})^2 = Var(\mu - \hat{\mu}) + (\mathbb{E}(\mu - \hat{mu}))^2 = Var(\mu - \hat{\mu}) = Var(\hat{\mu}) = \frac{\sigma^2}{n}$$

Otrzymujemy więc kolejno $MSE(0) = \frac{\sigma^2}{n} < 0.01$, czyli $n > 100\sigma^2 = 100\frac{\sum ln^2(x_i)}{n}$, skąd ostatecznie $n > 10\sqrt{\sum ln^2(x_i)}$.

□

Zadanie 21:
Na podstawie próby prostej $X_1, ..., X_n$ ze zmiennej losowej o skończonej wartości oczekiwanej μ, estymujemy μ za pomocą estymatora
$\hat{\mu} = \sum_{i=1}^{n} a_i X_i$,
gdzie $a_i > 0$ oraz $\sum_{i=1}^{n} a_i = 1$ Pokaż, że $\hat{\mu}$ jest estymatorem nieobciążonym. Dodatkowo zakładając, że istnieje wariancja X_1 znajdź $a_i, ..., a_n$ minimalizujące wariancję $\hat{\mu}$.

Rozwiązanie: Sprawdzamy czy estymator jest nieobciążony licząc jego wartość oczekiwaną: $\mathbf{E}(\hat{\mu}) = \mathbf{E}(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i \mathbf{E}(X_i)) = \sum_{i=1}^{n} a_i\mu = \mu\sum_{i=1}^{n} a_i = \mu$

Częśc B:$Var(\hat{\mu}) = \sum_{i=1}^{n} Var(a_i X_i) = \sum_{i=1}^{n} a_i^2 Var(X_i) = \sigma^2 * \sum_{i=1}^{n} a_i^2$

musimy zminimalizować sume a_i^2 niech $a_i = \frac{1}{n}$ zauważmy że $\frac{1}{n^2} + \frac{1}{n^2} \le \frac{2}{n^2} + a^2 + a^2 = (\frac{1}{n} - a)^2 + (\frac{1}{n} + a)^2$ zatem suma dla $a_i = \frac{1}{n}$ jest najmniejsza i wynosi $\frac{1}{n}\sigma^2$

Zadanie 22: Zadanie 22
Niech $X_1, ..., X_n$ będzie próbą prostą z rozkładu normalnego o wartości oczekiwanej μ i wairancji σ^2, gdzie μ jest nieznamnym parametrem, a σ^2 jest znana. Wyznacz a,b, takie że $[\overline{X} + a, \overline{X} + b]$ jest przedziałem ufności na poziomie $1 - \alpha$ o najmniejszej możliwej długości.

Rozwiązanie:
Zacznijmy od udowodnienia pomocniczego twierdzenia, które mówi, że dla rozkładu $N(0, 1)$ przedział o ustalonej długości będzie miał maskymalne prawdopodtheobieństwo wtedy tylko wtedy gdy gdy ten przedział jest symetryczny wokół zera.

Lemat 1:

Niech $a, b \in \mathbb{R}$, takie że $b - a = 2r \in \mathbb{R}_+$ dla ustalonego r. Wówczas $P(X \in [a, b])$ jest największe $\iff b = -a = r$, gdzie $X \approx N(0, 1)$.
Dowód
Zacznijmy od obserwacji że jeśli g jest funkcją gęstości zmiennej losowej X to g jest funkcją rosnącą na przedziale $(-\infty, 0)$, oraz funkcją malejącą na przedziale $(0, \infty)$. załóżmy najpierw że obie liczby a, b są po jednej stronie 0, (BSO $a, b > 0$).
Wówczas $\exists_{\epsilon>0}$, taki że $a - \epsilon > 0$. wtedy $\forall_{x\in[a,b]}\ g(x - e) > g(x)$ co implikuje, że $P(X \in [a - \epsilon, b - \epsilon]) > P(X \in [a, b])$, stąd takie a, b nie maksymalizują prawdopodobieństwa.
Pozostał nam przypadek gdy $a < 0, b > 0$. Załóżmy dodatkowo BSO ($|a| < |b|$, przypadek przeciwny jest analogiczny).
Wówczas $\exists_{\epsilon>0}$, taki że $|b - 2\epsilon| > |a|$. $P(X \in [a - \epsilon, b - \epsilon]) - P(X \in [a, b]) = P(X \in [a - \epsilon, a]) - P(X \in [b - \epsilon, b]) =$ (symetria rozkładu $N(0, 1)$) $= P(X \in [-a, -a - \epsilon]) - P(X \in [b - \epsilon, b]) > 0$ na mocy poprzedniego podpunktu, ponieważ oba przedziały leżą po jednej stronie i mają ta samą długość ϵ. Stąd dane a, b, także nie maskymalizują szukanego prawdopodobieństwa. To też dowodzi że przedział symetryczny będzie maskymalizował bo dowolne przesunięcie zmniejszy szukane prawdopodobieństwo.

Mając już dowód lematu możemy przejść do naszego problemu. Zauważmy, że skoro każdy X_i ma rozkład $N(\mu, \sigma^2)$ to $\overline{X}$ ma rozkład $N(\mu, \frac{\sigma^2}{n})$. Liczymy:
$1 - \alpha = P(\mu \in [\overline{X} + a, \overline{X} + b]) = P(\overline{X} + a \leq \mu \leq \overline{X} + b) = P(-a \geq \overline{X} - \mu \geq -b) = P(\frac{-b\sqrt{n}}{\sigma} \geq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{-a\sqrt{n}}{\sigma}) = \Phi(\frac{-a\sqrt{n}}{\sigma}) - \Phi(\frac{-b\sqrt{n}}{\sigma})$ gdzie Φ jest dystrybuantą standardowego rozkładu normalnego.
Stąd jeśli chcemy maksymalizować prawdopodobieństwo mając stałą długość przedziału (tzn. $b - a = const$) to na mocy lematu nasze a, b muszą być symetryczne względem zera. Po wstawieniu $a = -b$ otrzymujemy, że $\Phi(\frac{b\sqrt{n}}{\sigma}) = \frac{\alpha}{2}$, ostatecznie otrzymujemy $a = -b$ natomiast $b = \frac{\sigma z_{\frac{\alpha}{2}}}{\sqrt{n}}$

Zadanie 23:
Niech $X_1, ..., X_n$ będzie próbą prostą z rozkładu normalnego o wartości oczekiwanej μ i wariancji σ^2. Rozważmy estymator wariancji postaci:

$$S_c^2 = c\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Znajdź c, dla którego osiągamy najmniejszy błąd średniokwadratowy.

Rozwiązanie: Błąd średniokwadratowy jest równy sumie wariancji estymatora i kwadratu jego obciążenia. Dodatkowo, wiem, że $\frac{n}{c\sigma^2}c\sum_{i=1}^{n}(X_i - \bar{X})^2$ ma rozkład χ^2 z $n - 1$ stopniami swobody.

Zatem liczę:

$$Var(c\sum_{i=1}^{n}(X_i-\bar{X})^2)=2c^2\sigma^4(n-1)$$

$$E(c\sum_{i=1}^{n}(X_i-\bar{X})^2)=c\sigma^2(n-1)$$

Zatem błąd średniokwadratowy wynosi:

$$MSE=\sigma^4(c(n-1)-1)^2+2c^2\sigma^4(n-1)$$

. Po przekształceniach i policzeniu pochodnej po c, otrzymuję wielomian $c(n-1)-1+2c$, dla którego zero otrzymujemy w punkcie $c=\frac{1}{n+1}$

Zadanie 24:
Niech $X_1,...,X_{10}$ będzie próbą prostą z rozkładu normalnego o wartości oczewianej zero i wariancji σ^2. Skonstruj test najmocniejszy na poziomie istotności $\alpha=0.05$ do zweryfikowania hipotezy, że $\sigma^2=1$ przeciw hipotezie $\sigma^2=\sigma_1^2$, dla $\sigma_1^2>1$. Dla jakich σ_1^2 moc tego testu będzie większa od 0.95? Wskazówka: kwantyle rozkładu $\chi^2(0.95,10)=18.30$, $\chi^2(0.05,10)=3.94$

Rozwiązanie: Wiemy, że dla pewnej statystyki X mamy dane α oraz $1-\beta$ (czyli moc) następującymi wzorami:

$$\alpha:\mathbb{P}(X\geq c|H_0)$$

$$1-\beta:\mathbb{P}(X\geq c|H_1)$$

gdzie c to pewien obszar krytyczny. Nazwijmy naszą statystykę Z_n i rozpiszmy dla niej powyższe:

$$\mathbb{P}(Z_n\geq c_1|H_0)=0.05$$

$$\mathbb{P}(Z_n\geq c_2|H_1)=0.95$$

mamy stąd:

$$c_1=\chi^2(0.95,9)$$

$$c_2=\chi^2(0.05,9)$$

(9, bo $df=n-1$, gdzie $n=10$). Następnie mamy więc:

$$\mathbb{P}(\frac{nS_n^2}{\sigma_0^2}\geq c_1|H_0)=0.05$$

$$\mathbb{P}(\frac{nS_n^2}{\sigma_1^2}\geq c_2|H_1)\geq 0.95$$

Z tego mamy:

$$\mathbb{P}(S_n^2\geq\frac{c_1\sigma_0^2}{n}|H_0)=0.05$$

$$\mathbb{P}(S_n^2\geq\frac{c_2\sigma_1^2}{n}|H_1)\geq 0.95$$

Prawe strony tych nierówności muszą być takie same, ponieważ to ten sam obszar krytyczny. Mamy stąd:

$$\frac{c_1\sigma_0^2}{n}=\frac{c_2\sigma_1^2}{n}$$

$$\sigma_1^2=\frac{c_1}{c_2}$$

$$\sigma_1^2 = \frac{\chi^2(0.95, 9)}{\chi^2(0.05, 9)}$$

Mimo że w treści zadania zostały podane złe kwantyle, bo $df = n - 1$, a $n = 10$, więc $df = 9$ a nie $df = 10$, skorzystam z nich

$$\sigma_1^2 = \frac{18.30}{3.94}$$

$$\sigma_1^2 = 4.65$$

Więc, aby moc testu była większa od 0.95 to $\sigma_1^2 > 4.65$.