Laboratiorium 2 ze statystyki

Cele

- Celem zadań 1,2 i 4 jest zastosowanie twierdzenia z wykładu mówiącego, że jeśli $\hat{\theta}$ jest ENW parametru θ a $g = g(\theta)$ funkcją parametru θ, to $g(\hat{\theta})$ jest ENW wielkości g i porównanie wyników z oczekiwaniami.
- W zadaniu 4 użyjemy wykresu kwantylowego do sprawdzenia czy rozkład danych jest normalny.
- Zadanie 5 to porównanie wyników eksperymentalnego wyznaczania ENW dla rozkładów: normalnego i Laplace'a.
- W zadaniu 6 sprawdzamy, jaki wpływ na szacowania ma rozmiar próby.

Każde z zadań zawiera krótkie podsumowanie.

Zadanie 1 - rozkład dwumianowy

Wielkość $P(X \leq 3)$ jest funkcją parametru p:

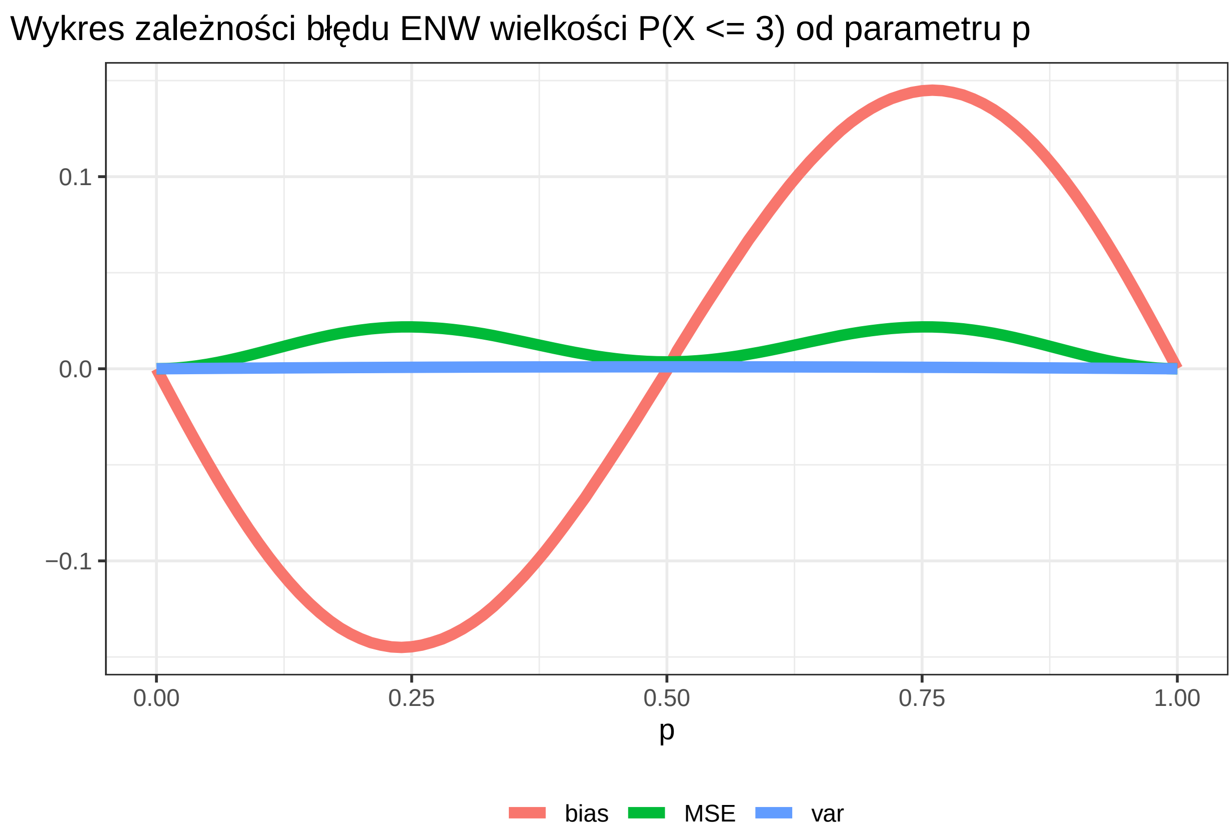
$$P(p) = \binom{5}{3} p^3 (1-p)^{1-3} + \binom{5}{4} p^4 (1-p)^{1-4} + \binom{5}{5} p^5 (1-p)^{1-5}$$

stąd estymator nawiększej wiarogodności P otrzymamy wyznaczając ENW p (dla tego rozkładu jest nim proporcja sukcesów) i przekazując otrzymaną wartość do funkcji $P(p)$.

Table 1: Średni błąd kwadratowy, wariancja i obciążenie estymatora największej wiarogodności P dla wybranych wartości parametru

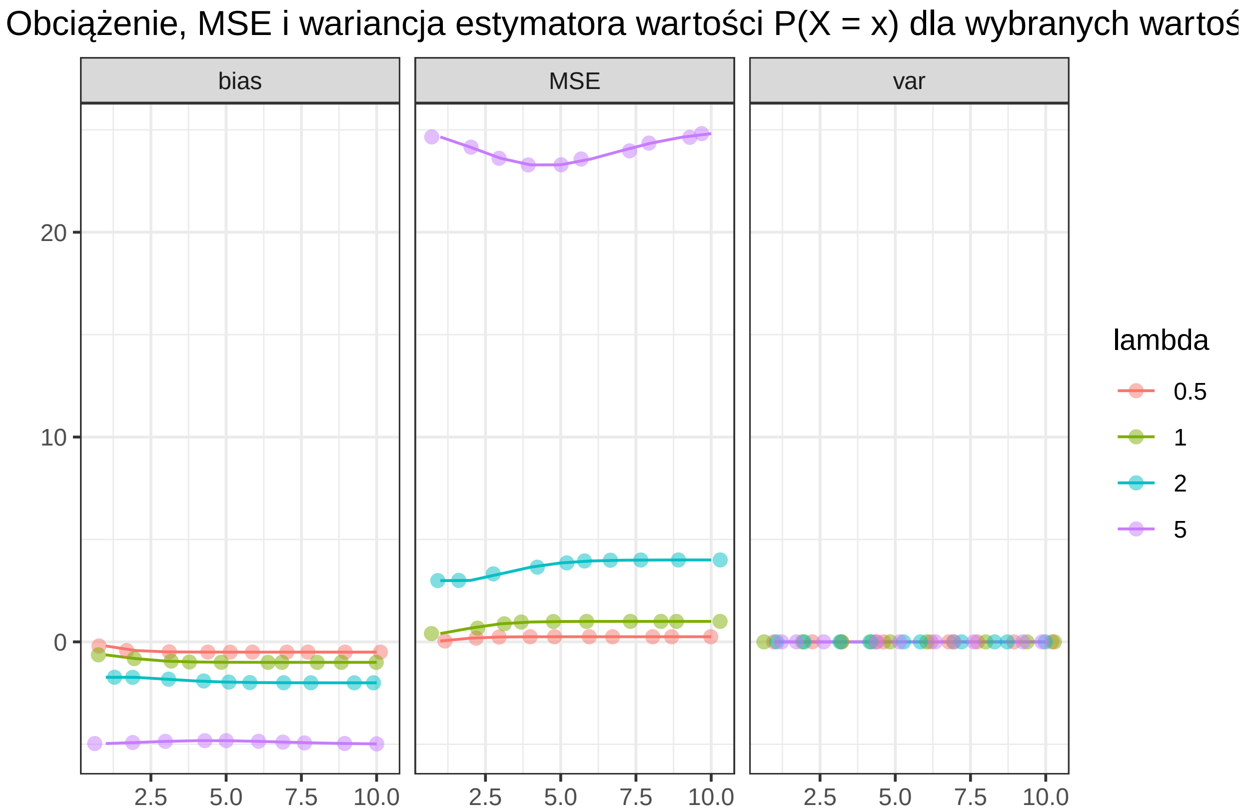
p	n	MSE	bias	var
0.1	50	0.0082477	-0.0906803	0.0003642
0.3	50	0.0198273	-0.1353981	0.0008587
0.5	50	0.0035856	-0.0008535	0.0010366
0.9	50	0.0082477	0.0906803	0.0003642

1. Estymowane wartości $P(X \leq 3)$ są bliskie prawdziwym teoretycznym wartościom, a MSE, obciążenie i wariancja małe - oznacza to, że wybrany sposób estymacji jest bardzo dobry.
2. Wykonując dodatkowe eksperymenty można zauważyć, że wartości MSE, obciążenia i wariancji są symetryczne względem $p = 0.5$.

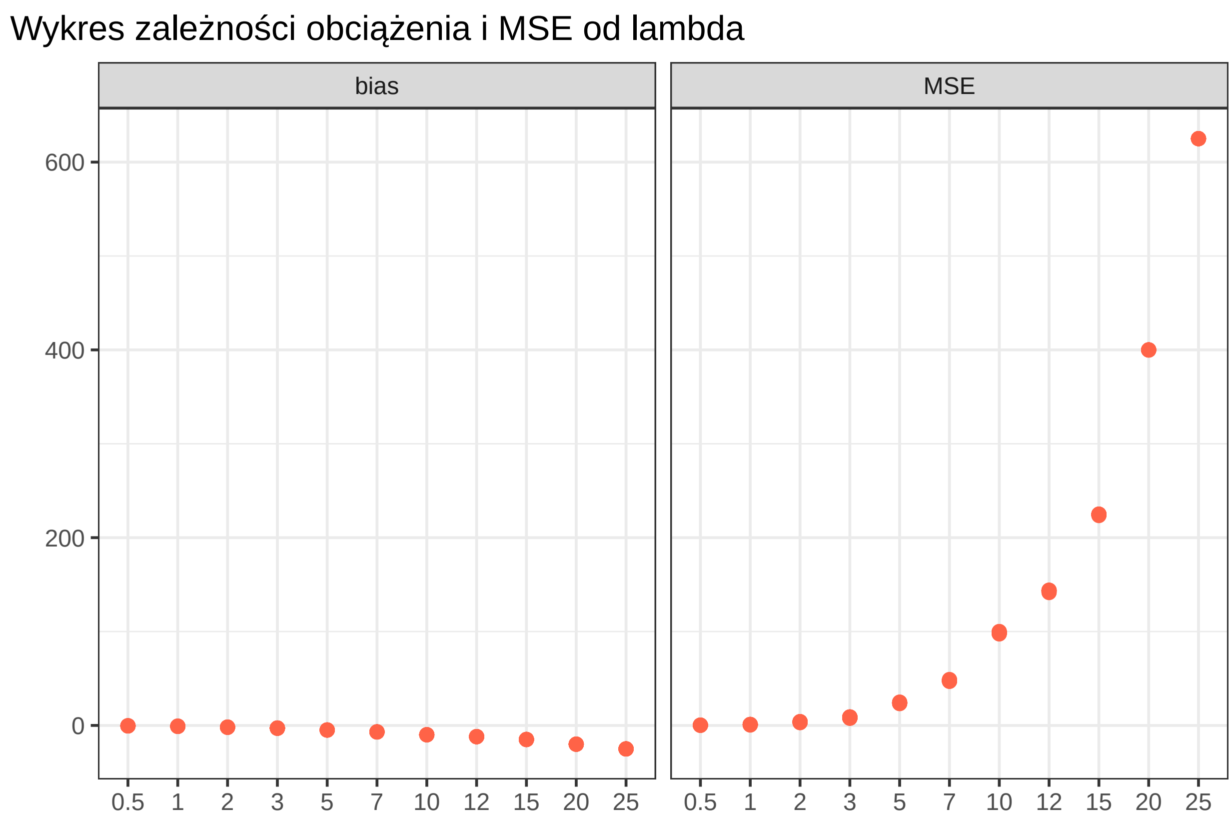


Zadanie 2 - rozkład Poissona

Estymatorem największej wiarogodności parametru λ jest średnia arytmetyczna. W celu wyznaczenia ENW $P(X = x)$, korzystamy z tego samego sposobu co w zadaniu 1.



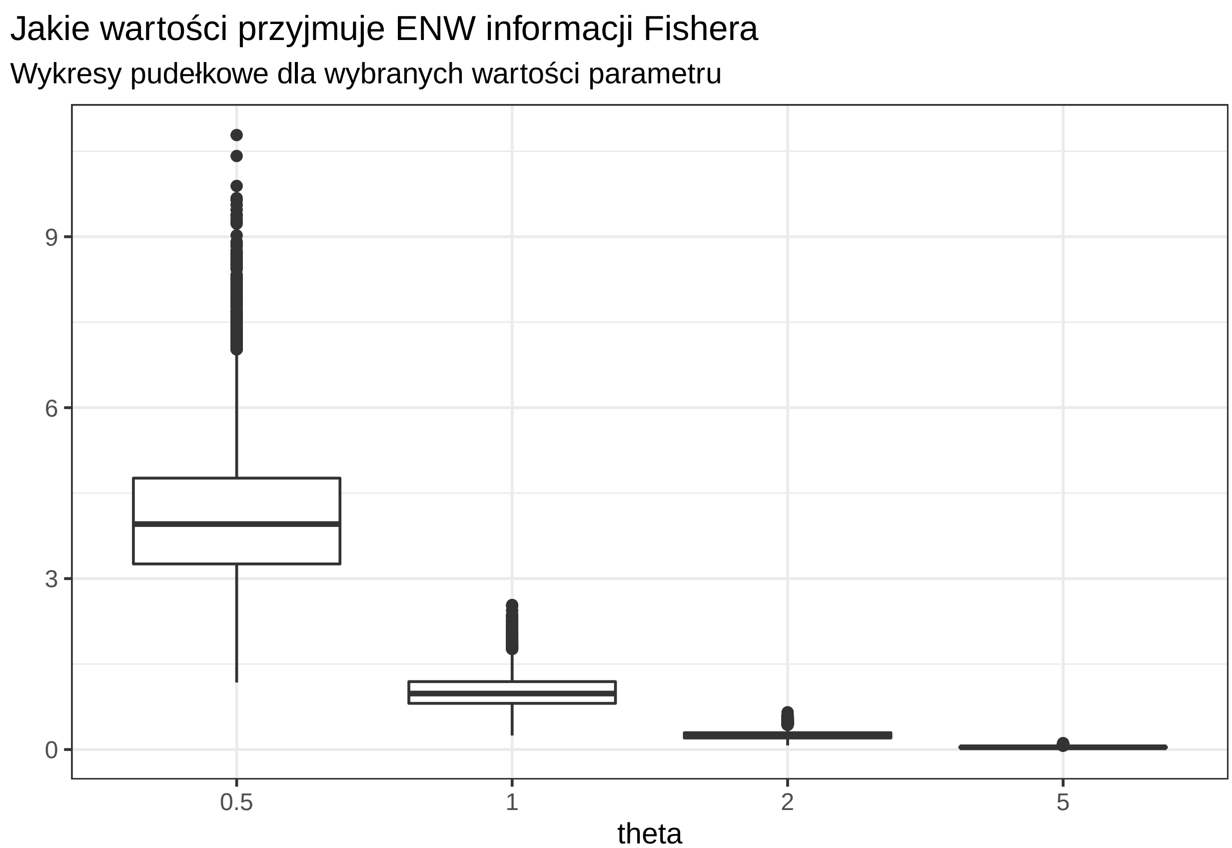
1. Wariancja jest niezależnie od λ praktycznie zerowa.
2. Obciążenie i bezwzględna wartość średniego błędu kwadratowego rosną wraz ze wzrostem λ, ale nie różnią się znacząco dla różnych wartości x. Przeprowadziłam dodatkowe eksperymenty i sprawdziłam jak wygląda zależność obciążenia i MSE od λ.



Zadanie 4

Gdy $\beta = 1$, to: funkcja masy prawdopodobieństwa jest postaci $P(x|\theta) = \theta x^{\theta-1}$, informacja Fishera jest funkcją parametru θ: $I(\theta) = \frac{1}{\theta^2}$, a estymator największej wiarogodności θ można wyrazić zwartym wzorem: $\hat{\theta} = \frac{-n}{\sum ln(x_i)}$. Analogicznie do poprzednich zadań $\hat{I(\theta)} = I(\hat{\theta})$.

W pierwszej cześci zadania estymujemy informację Fishera dla wartości θ uzyskanych dla 10000 różnych prób - na wykresie widać, że estymatory przyjmują wartości zbliżone do takich, jakich możnaby się spodziewać, tzn. że ich średnie wartości są odwrotnie proporcjonalne do λ^2. Ponadto, większe wartości parametru skutkują większą wariancją ENW informacji.



W dalszej części za zapamiętaną wartość estymatora $I(\hat{\theta})$ przyjmiemy średnią arytmetyczną wyników - takie oszacowanie powinno mieć tym mniejsze znaczenie, im większa będzie wartość θ ze względu na mniejszy rozrzut estymatora. Uwagi :

1. prezentowane wyniki są obliczane dla innych prób losowych niż użyte wcześniej.
2. krzywa normalna naniesiona na histogramy rysowana jest z użyciem średniej i odchylenia standardowego liczonych dla wszystkich wartości Y (tzn. $\theta = 0.5, 1, 2, 5$). Nie robi to dużej różnicy, gdyż indywidualne wyniki są pbardzo podobne.

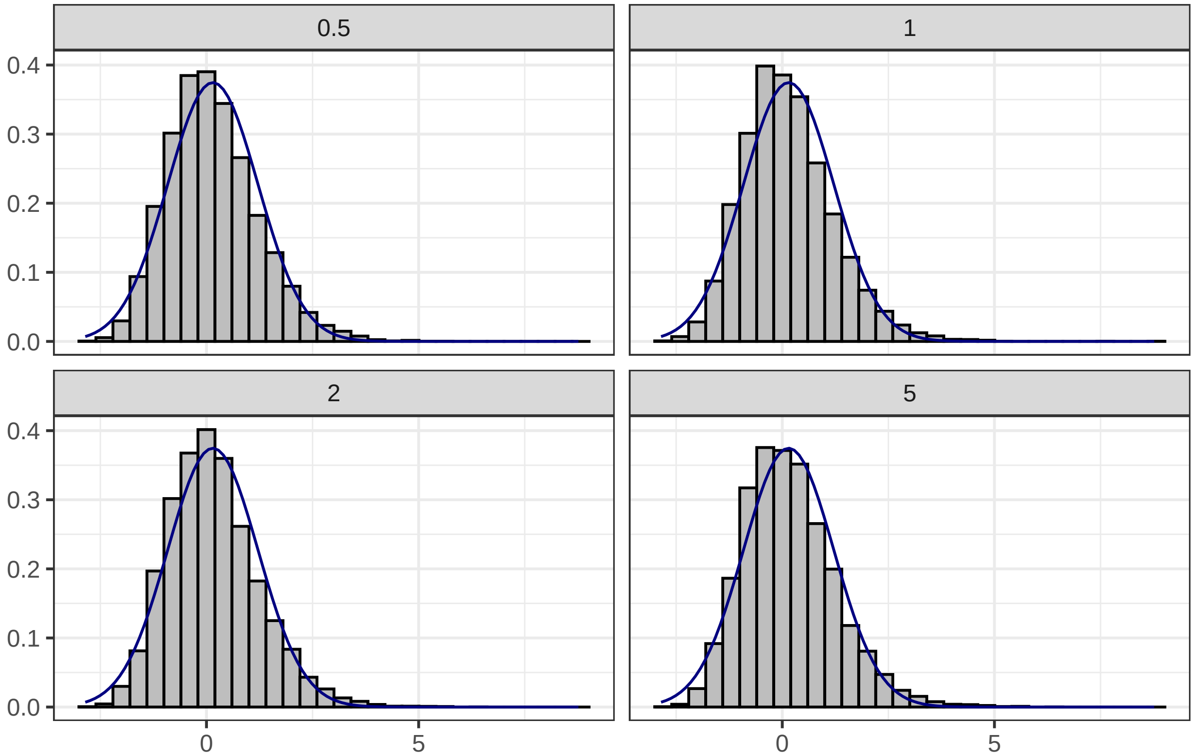
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

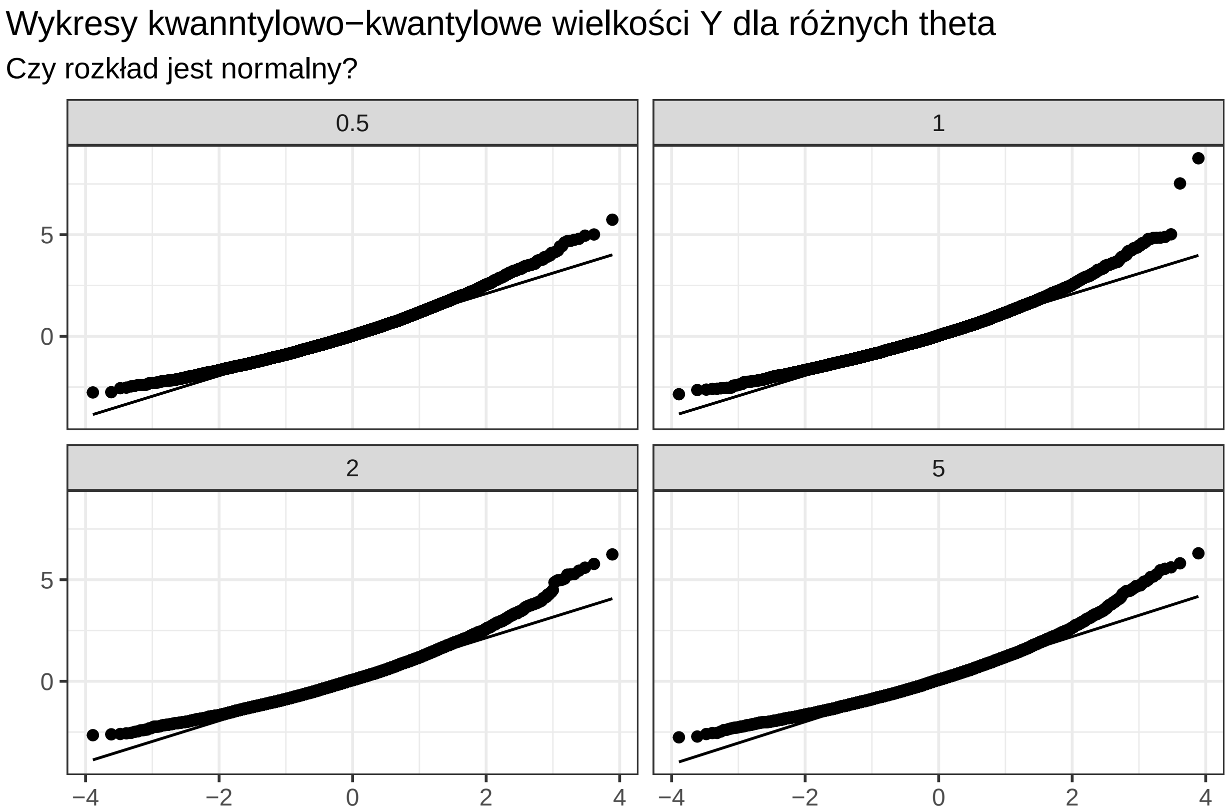
Table 2: Porównanie średniej i wariancji Y dla różnych wartości parametru

th	mean	sd
0.5	0.1366671	1.054771
1	0.1363110	1.060381
2	0.1597466	1.063109
5	0.1691760	1.080923

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Histogramy gęstości uzyskanych wartości Y dla różnych wartości theta; n = 50





Wygląd wykresów pozwala stwierdzić, że rozkład wielkości Y jest bardzo bliski rozkładowi normalnemu i lekko skośny względem niego.

Zadanie 5

Table 3: Wyniki zadania 5c

	MSE	bias	var
th1	0.0395111	-0.0015393	0.0395126
th2	0.0240074	-0.0010795	0.0240087
th3	2.7499850	0.0304426	2.7493332
th4	0.0399686	0.1025496	0.0294551

Przypadku wszystkich testowanych wartości parametrów najlepszym estymatorem okazał się θ_2, czyli mediana. Nie ma w tym nic dziwnego - poprzednio była nim średnia, która jest ENW parametru dla rozkładu normalnego, a mediana jest ENW parametru dla rozkładu Laplace'a.

Zadanie 6.1

Po wykonaniu eksperymentu dla $n = 20$ i $n = 100$ okazało się, że rozmiar próby nie ma wpływu na kształt wykresu przedstawionego w zadaniu 1. Dla jeszcze mniejszych prób można zaobserwować różnice, co można wyjaśnić tym, że $n = 20$ często przyjmuje się za minimalną do uzyskania zgodnych z teorią wyników liczebność próby.

Zadanie 6.2

Zmiana liczebności próby nie wpływa znacząco na zależności wariancji, średniego błędu kwadratowego i obciążenia od parametru, natomiast ich bezwzględne wartości maleją wraz ze wzrotem n.

Zadanie 6.4

Eksperymenty wykazały, że dla większej liczebności próby punkty na wykresie kwantylowo-kwantylowym leżą bliżej prostej, co oznacza że rozkład jest bardziej zbliżony do normalnego. Można to uzasadnić tym, że Y dąży w rozkładzie do rozkładu normalnego.

Zadanie 6.5

Wyniki dla $n = 100$ są zbliżone do wyników dla $n = 50$, z trochę mniejszymi wartościami błędów dla sensownych estymatorów. W przypadku $n = 20$ widać większą losowość - znów ma to związek ze zbyt małą do uzyskania zgodnych z teorią liczebnością.