Analiza przeżycia
Raport 1

Romana Żmuda

9 listopada 2020

Spis treści

1 Lista 1 **3**

2 Tworzenie wykresów funkcji **4**

3 Statystyki opisowe i graficzna prezentacja danych **5**

4 Zadania do sprawozdania 1, część 1 **15**
4.1 zabawa 15

5 Zadanie do sprawozdania - Część 1 **16**
5.1 Zadanie 1 16

1 Lista 1

W pakiecie R standardowo dostępne są funkcje gęstości, dystrybuanty i kwantyli podstawowych rozkładów takich jak: beta (beta), dwumianowy (binom), Cauchy'ego (cauchy), chi-kwadrat (chisq), wykładniczy (exp), Fishera (f), gamma (gamma), geometryczny (geom), hypergeometryczny (hyper), logarytmiczno-normalny (lnorm), logistyczny (logis), ujemny dwumianowy (nbinom), normalny (norm), Poissona (pois), Studenta (t), jednostajny (unif), Weibulla (weibull), Wilcoxona (wilcox) (zobacz Paradis, 2005, str. 17).

Poprzedzając nazwę rozkładu literą d (od ang. *density*) uzyskujemy funkcję gęstości rozkładu; literą p (od ang. *probability*) - dystrybuanę; literą q (od ang. *quantile*) - funkcję kwantylową.

W pliku raportu komendy kodu R piszemy w tzw. wstawce (ang. *chunk*). W tzw. *chunk*-u, między <<>>= możemy wpisać jego nazwę i opcje (lokalne dla danego *chunk*-u), np. echo=TRUE - pokazuje komendę wykonania, FALSE - ukrywa ją, fig=TRUE - przekazuje wykres do umieszczenia - FALSE - ukrywa wykres, eval ustala czy wyniki są obliczane. Jezeli ponadto nie chcemy, by wynik wykonania fragmentu kodu był wpisywany do pliku wynikowego, to za argumenty wstawki należy podać results=hide.

Na przykład pisząc

```
> dexp(1,rate=2)

[1] 0.2706706
```

uzyskujemy wartość gęstości rozkładu wykładniczego o średniej 1/2, w punkcie 1.

Natomiast, po wpisaniu

```
> pexp(1,rate=2)

[1] 0.8646647

> qexp(0.5,rate=2)

[1] 0.3465736
```

otrzymujemy odpowiednio wartość dystrybuanty rozkładu wykładniczego o średniej 1/2 w punkcie 1 i wartość kwantyla rzędu 0.5 rozkładu wykładniczego o średniej 1/2.

Liczby losowe (pseudolosowe) z danego rozkładu możemy uzyskać poprzedzając nazwę rozkładu literą r. Na przykład po wpisaniu

```
> rexp(10,rate=2)

 [1] 0.003803275 0.019355135 0.099979826 0.107211888 0.192717337 0.862287357
 [7] 1.123722977 1.305180817 0.088252806 0.050596705
```

uzyskamy 10 liczb wygenerowanych z rozkładu wykładniczego o średniej 1/2.

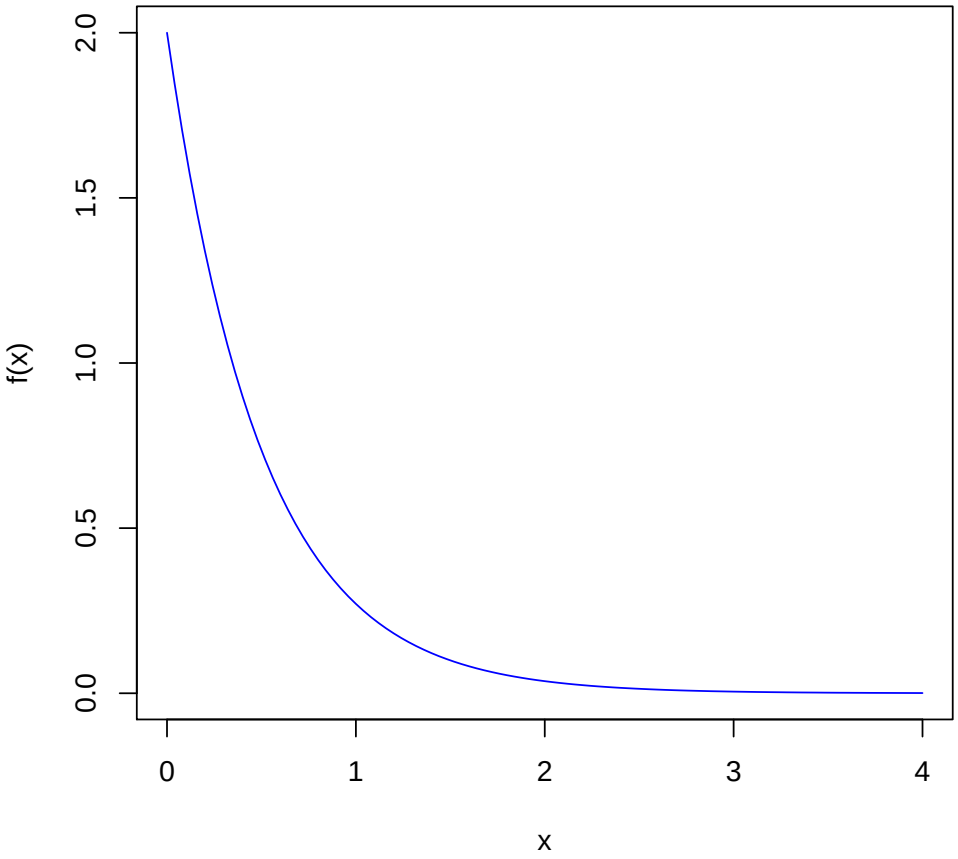
Znając postać funkcji odwrotnej F^{-1} do dystrybuanty F rozkładu absolutnie ciągłego względem miary Lebesgue'a, możemy wygenerować dane z tego rozkładu, korzystając z metody dystrybuanty odwrotnej. Mianowicie, można pokazać, że jeżeli $U_1, \ldots, U_n$ są zmiennymi losowymi z rozkładu jednostajnego na odcinku $(0, 1)$, to zmienne losowe $F^{-1}(U_1), \ldots, F^{-1}(U_n)$ są zmiennymi losowymi z rozkładu o dystrybuancie F.

2 Tworzenie wykresów funkcji

Wykresy funkcji możemy łatwo uzyskać korzystając z funkcji *curve*. Na przykład wykres gęstości rozkładu wykładniczego o średniej 1/2 (zobacz rysunek 1) otrzymamy w wyniku następującego kodu.

```
> curve(dexp(x,rate=2),xlim=c(0,4),lty=1,xlab='x',ylab='f(x)',col='blue')
```

Powyżej *lty* oznacza rodzaj linii. I tak: 1, to ciagła (domyślna), 2 — kreskowana, 3 — kropkowana, 4 — kropka-kreska, 5 — długa kreska, 6 — podwójna kreska. Natomiast *xlab* i *ylab* oznaczają podpisy odpowiednio osi X i Y; *main* - tytuł wykresu, *xlim* - zakres osi X. Więcej opcji opisanych jest np. w rozdziale 4.2 Paradisa (2005).



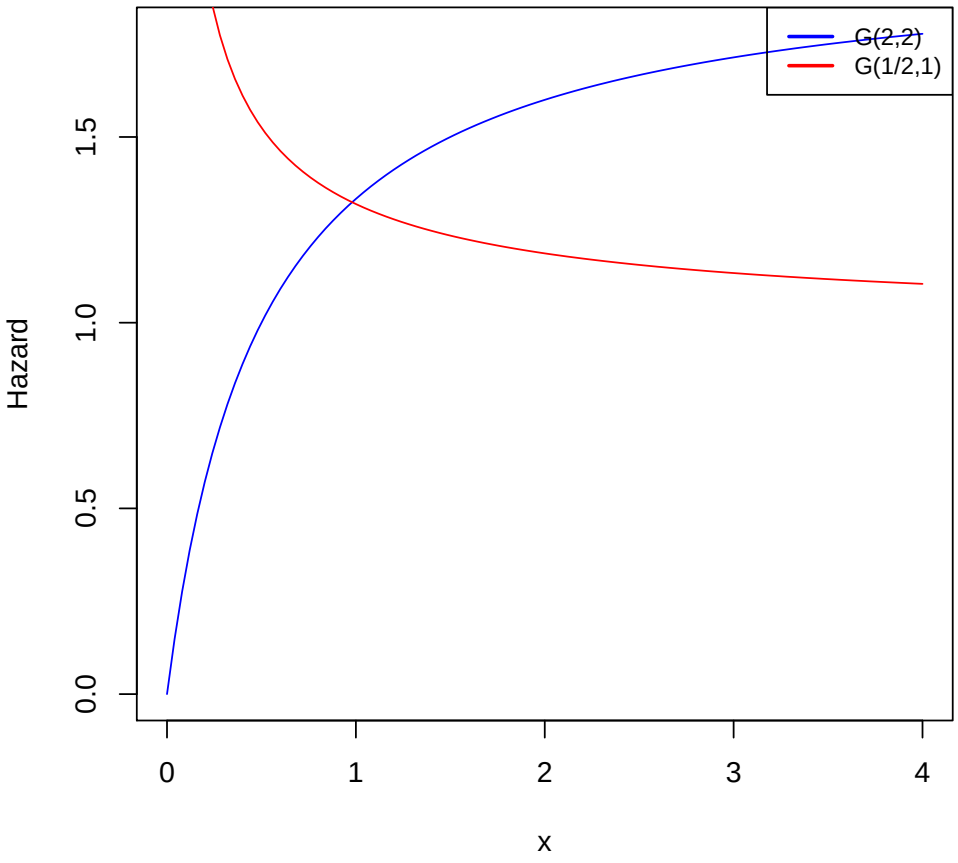
Rysunek 1: Wykres gęstości rozkładu wykładniczego $\mathcal{E}(2)$

W celu utworzenia wykresów funkcji hazardu, np. rozkładu $\mathcal{G}(2,2)$ i rozkładu gamma $\mathcal{G}(2,1)$) na jednym rysunku, możemy (ale nie musimy) najpierw zdefiniować funkcje hazardu rozkładu gamma.

```
> hgamma<-function(x,alfa,beta)
+ {
+   return(dgamma(x,alfa,beta)/(1-pgamma(x,alfa,beta)))
+   }
```

Wykresy funkcji hazardu rozkładów gamma $\mathcal{G}(2,2)$, $\mathcal{G}(1/2,1)$ (zobacz rysunek 2) możemy uzyskać następująco.

```
> curve(dgamma(x,2,2)/(1-pgamma(x,2,2)),xlim=c(0,4),col='blue',ylab='Hazard');
> curve(dgamma(x,1/2,1)/(1-pgamma(x,1/2,1)),col='red',add=T)
> legend("topright", c("G(2,2)","G(1/2,1)"), col=c("blue","red"),
+        lwd=2, cex=.85 )
```



Rysunek 2: Wykresy funkcji hazardu rozkładu gamma $\mathcal{G}(2,2)$ i $\mathcal{G}(1/2,2)$

3 Statystyki opisowe i graficzna prezentacja danych

Niech x i y będą realizacjami niezależnych prób rozmiaru 100 z rozkładu wykładniczego odpowiednio $\mathcal{E}(1)$ i $\mathcal{E}(5)$.

```
> set.seed(123456)
> x<-rexp(100,1)
> x
```

```
  [1] 0.59556864 0.50713019 1.25816992 1.05937397 1.13832363 1.97305224
  [7] 0.06971592 2.62386140 0.11615031 0.61055776 0.13696987 1.03312158
```

```
 [13] 0.55441263 1.66304280 2.36976555 1.44749707 1.76153755 1.17870009
 [19] 0.11505378 0.04593751 0.08254059 0.64790245 2.36387322 0.49362741
 [25] 0.21080902 1.89894264 0.87100014 0.03070503 1.09182126 0.35634527
 [31] 1.14121193 1.03137701 0.49771200 0.01323310 0.52648497 0.66887447
 [37] 0.67407386 0.76301554 1.79375431 1.95457977 3.71886799 0.05545480
 [43] 1.83933093 1.56817234 0.18194804 6.07609318 1.13721158 1.42043595
 [49] 4.29365204 0.24007576 1.42136266 0.09165579 0.24290944 1.52242452
 [55] 6.58125637 0.66686619 0.27843291 1.86677074 0.09508150 0.12577113
 [61] 1.67403007 0.92455470 1.13126898 1.72262600 1.93370181 0.72049716
 [67] 0.53631889 2.10616492 0.31889504 0.93902129 1.90284755 0.25214263
 [73] 0.03343692 0.04046412 0.23866960 1.16151016 2.87007288 2.16762685
 [79] 0.97945646 1.29407512 1.94991192 0.87210015 2.06514119 0.24577688
 [85] 0.13955904 4.09864896 0.45177480 0.82295080 1.30461279 0.19227490
 [91] 1.04612364 0.91076889 1.42817604 0.32343915 0.42758572 0.60931564
 [97] 2.90967265 1.15716048 2.27591788 0.03637644
```

```
> y<-rexp(100,5)
```

Wartości podstawowych statystyk opisowych możemy uzyskać poprzez wywołanie funkcji *summary*.

```
> summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01323 0.30878 0.93179 1.17008 1.68618 6.58126
```

```
> summary(y)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.006991 0.067729 0.135424 0.188514 0.254159 0.926944
```

Wartości próbkowych odchyleń standardowych uzyskamy dzięki funkcji *sd*.

```
> sd(x)
```

```
[1] 1.169046
```

```
> sd(y)
```

```
[1] 0.1791627
```

Wartości statystyk opisowych możemy również przedstawić w ładniejszej tabeli. Mianowicie, korzystając z poniższego polecenia, otrzymujemy tabelę **??** z wartościami podstawowych statystyk opisowych obliczonych dla wektora x.

```
> #stat.opis<-c(summary(x),sd(x))
> #stat.opis.m=matrix(stat.opis,nrow=1,ncol=length(stat.opis), list(c(" "),c("Min","1s
>  #                  "Średnia","3d Qu.","Max","Odch.Stand.")))
> #print(xtable(stat.opis.m,caption="Statystyki opisowe x",label="tab:stx"), type="lat
```

Analogicznie, wartości statystyk opisowych dla wektora y zawarte są w tabeli 1.

```
> stat.opis.y<-c(summary(y),sd(y))
> stat.opis.y.m=matrix(stat.opis.y,ncol=length(stat.opis.y),nrow=1,,
+                     list(c(" "),c("Min","1st Qu","Mediana",
+                                   "Średnia","3d Qu.","Max","Odch.Stand.")))
> print(xtable(stat.opis.y.m,caption="Statystyki opisowe y",label="tab:sty"),
+       type="latex", table.placement = "H",,caption.placement = "top")
```

Tabela 1: Statystyki opisowe y

Min	1st Qu	Mediana	Średnia	3d Qu.	Max	Odch.Stand.
0.01	0.07	0.14	0.19	0.25	0.93	0.18

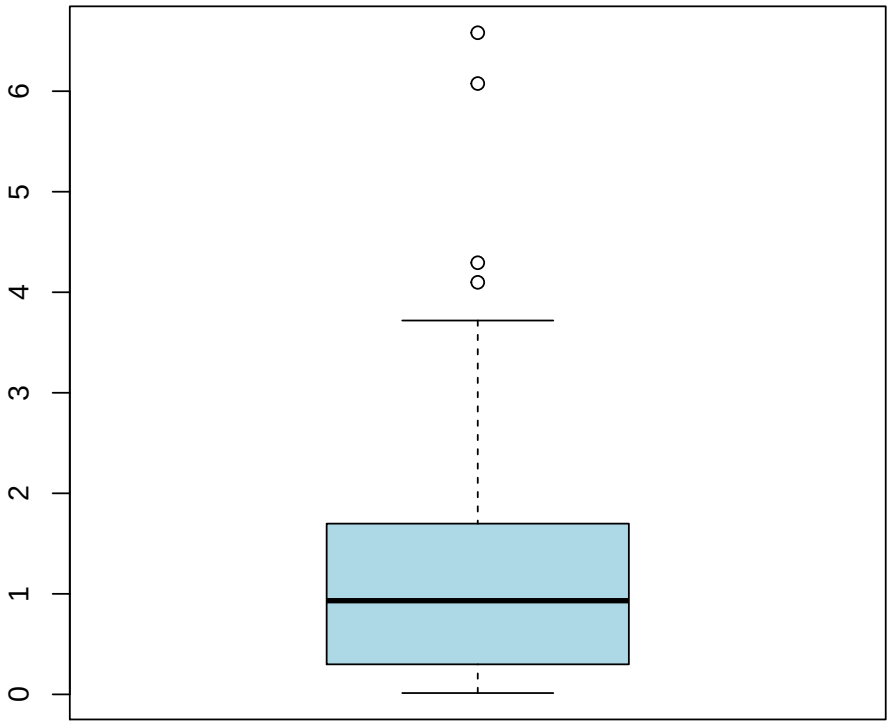
Jezyk R zawiera różne funkcje operujące na wektorach. Poniżej lista najważniejszych z opisem ich wartości (zobacz również Paradis, 2005, str, 32-33).

- max — maksymalna wartość z wektora,
- min — minimalna wartość z wektora,
- mean - średnia arytmetyczna (jeśli podamy dodatkowy parametr trim, to funkcja policzy średnią po odrzuceniu okreslonego odsetka wartosci skrajnych, np. mean(x,trim=0.1), to średnia z x po odrzuceniu 10% wartości skrajnych),
- median — mediana,
- quantile - dowolny kwantyl, np. quantile($x, .5$), to mediana z x,
- sd — odchylenie standardowe (pierwiastek z wariancji nieobciążonej),
- var - wariancja (nieobciążona),
- length — długosc wektora (liczba jego elementów),
- sum - suma elementów wektora (W przypadku wektorów logicznych sumę stanowi liczba elementów o wartosci TRUE),
- prod — iloczyn elementów,
- sort - wektor z wartościami uporzadkowanymi rosnąco (wektor statystyk pozycyjnych),
- pmin, pmax - funkcje te operują na kilku wektorach tej samej długosci; wartością tych funkcji jest wektor zawierający odpowiednio najmniejsze lub największe elementy wybrane z poszczególnych wektorów,
- cummin, cummax — funkcje te zwracają wektory zawierajace dla każdego elementu wartości odpowiednio minimalne, maksymalne znalezione „dotychczas”, czyli od pierwszego elementu do aktualnego,
- which - wektor zawierajacy indeksy, przy których argument ma wartosc TRUE,
- diff - wektor krótszy o 1, zawierajacy różnice między sasiadującymi elementami,
- rank - wektor rang .

W celu ilustracji danych najczęściej tworzy się histogramy i wykresy pudełkowe. Wykres pudełkowy dla wektora x, uzyskany np. poleceniem

```
> boxplot(x,col="lightblue",bg="yellow",main="Wykres pude■lkowy x")
```

zamieszczony jest na rysunku 3.

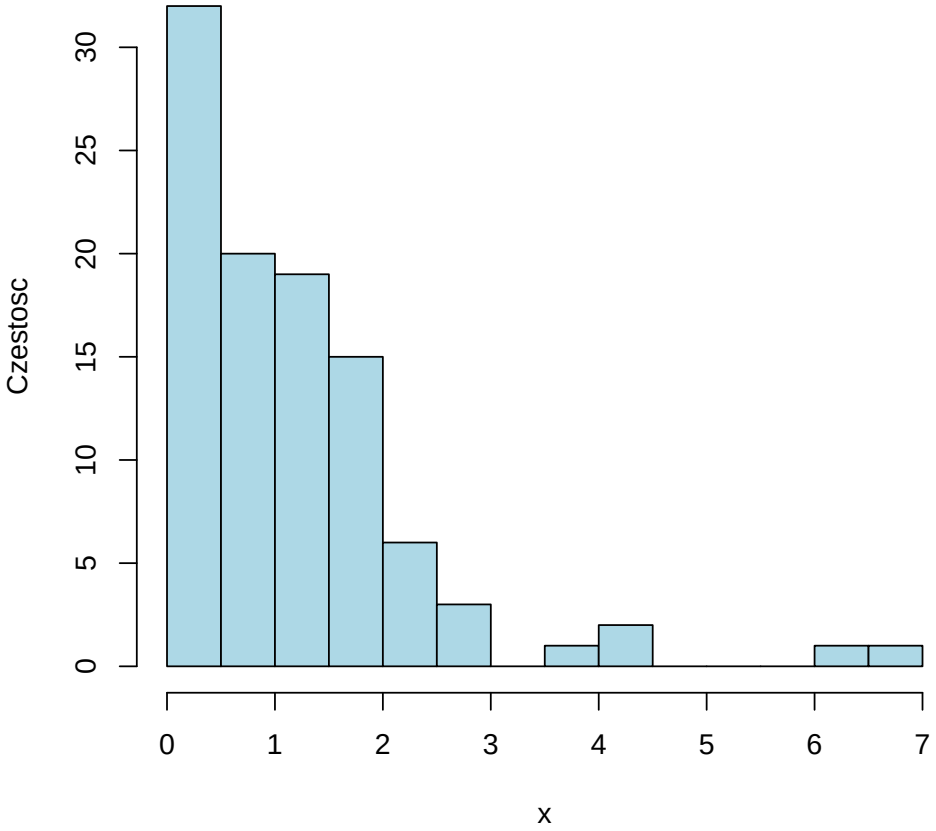


Rysunek 3: Wykres pudełkowy dla wektora obserwacji x

Histogramy możemy narysować przyjmując ustaloną liczbą przedziałów klasowych (breaks), równą np. 10, lub wybrać liczbę przedziałów klasowych zgodną z pewnymi regułami (Sturgesa, Scotta, czy Freedmana-Diaconisa).

Rysunek 4 przedstawia histgram uzyskany na podstawie następującego polecenia.

```
> hist(x, breaks=10, probability=FALSE, ylab="Częstość", main='',
+      col="lightblue")
```

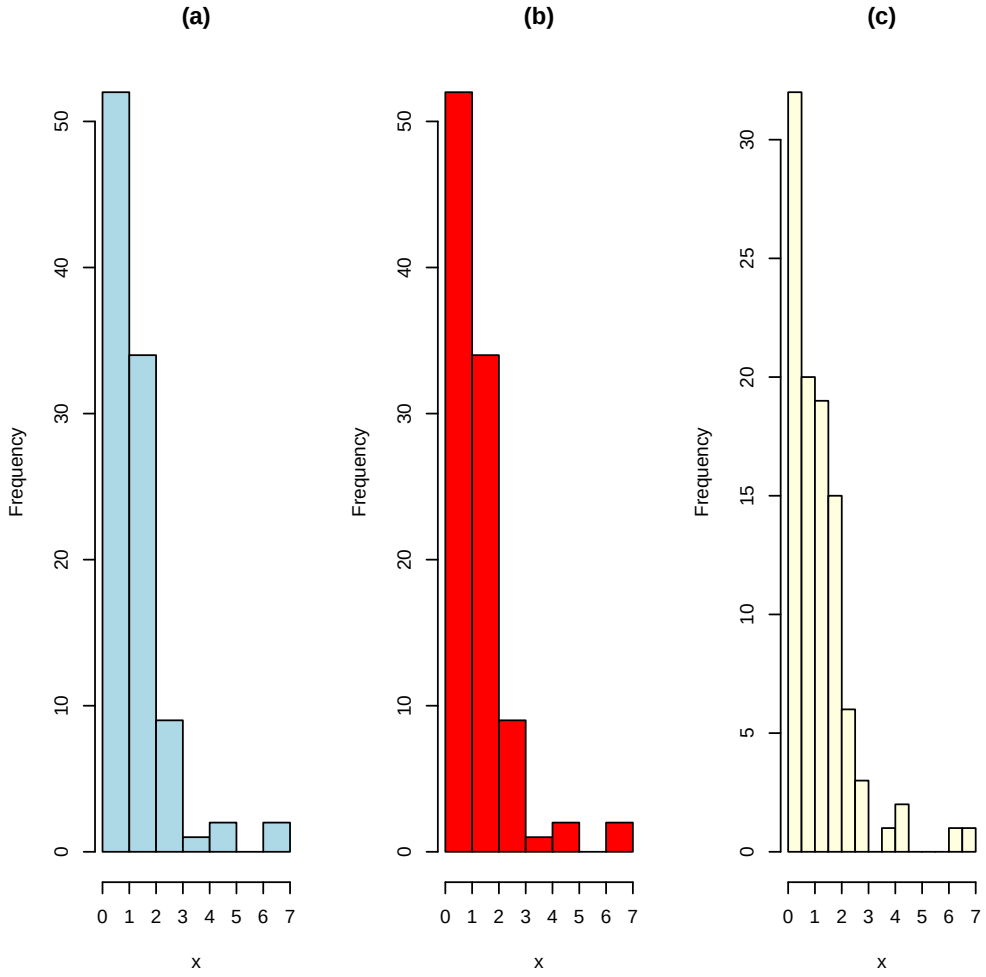


Rysunek 4: Histogram z ustaloną liczbą przedziałów klasowych, odpowiadający wektorowi x

Możemy również, korzystając z poniższego polecenia, narysować histogramy z liczbą przedziałów klasowych obliczonych zgodnie z regułami, dostępnymi w pakiecie.

```
> par(mfrow=c(1,3))
> hist(x, breaks = "Sturges", main="(a)", col="lightblue")
> hist(x, breaks = "Scott", main="(b)", col="red")
> hist(x, breaks = "Freedman-Diaconis",
+      main="(c)", col="lightyellow")
```

Wynik powyższego polecenia zamieszczony jest na rysunku 5.



Rysunek 5: Histogramy z liczbą przedziałów klasowych obliczonych zgodnie z regułą (a) Sturgesa, (b) Scotta, (c) Freedmana-Diaconisa, odpowiadające wektorowi x

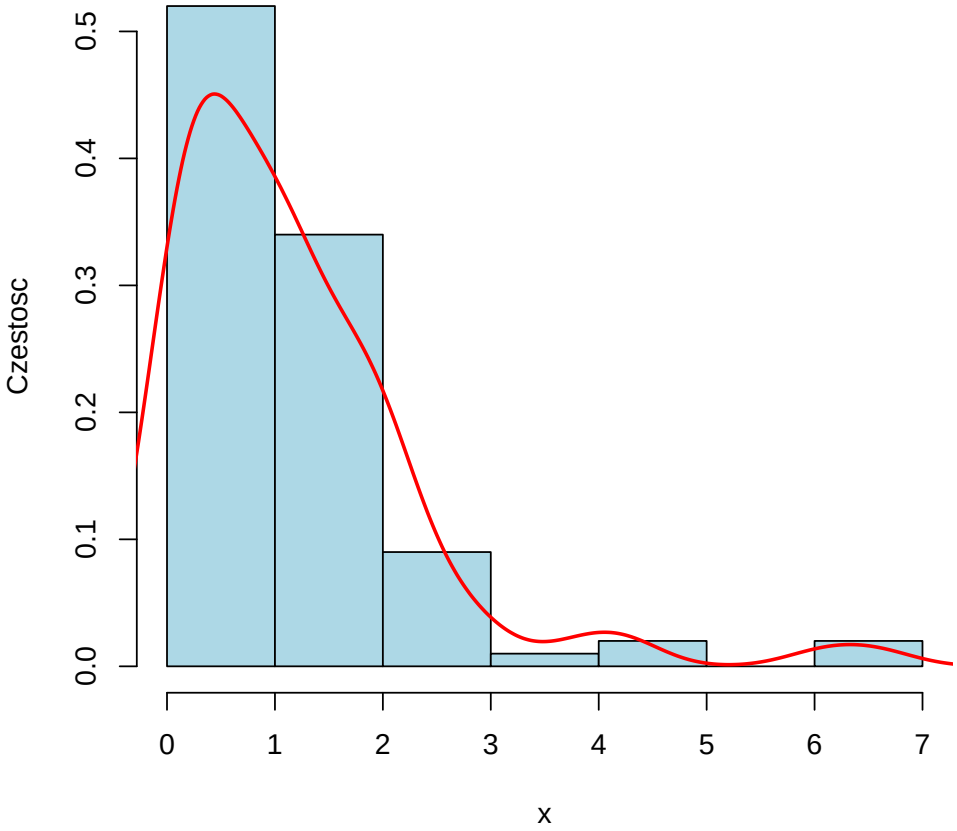
Na rysunku 6 znajduje się histogram z liczbą przedziałów klasowych obliczonych przy użyciu reguły Scotta i dodatkowo z narysowanym wykresem estymatora jądrowego gęstości, odpowiadający danym x, utworzony, przy wykorzystaniu następującego polecenia.

```
> h<-hist(x, breaks="Scott", probability=TRUE, xlim=c(0,7), ylab="Częstość",
+         col="lightblue")
> d<- density(x)
> lines(d, col="red", lwd=2)
```

Dla przypomnienia, estymator jądrowy gęstości wyraża się wzorem

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

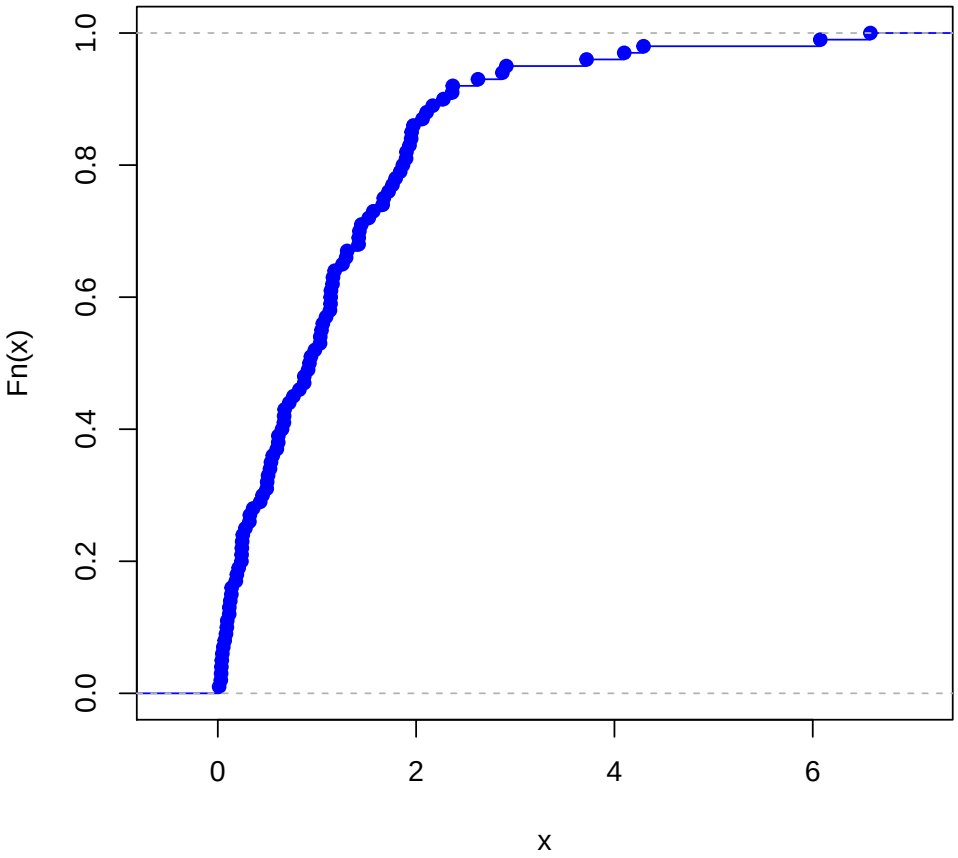
gdzie K jest funkcją jądrową, h - szerokością okna.



Rysunek 6: Histogram i wykres estymatora jądrowego gęstości odpowiadający danym x

W analizie przeżycia bardziej interesują nas wykresy estymatora funkcji przeżycia i funkcji hazardu niż histigramy. Najprostszy estymator funkcji przeżycia jest postaci $\hat{S}(t) = 1 - \hat{F}_n(t)$, gdzie $\hat{F}_n$ jest dystrybuantą empiryczną. Wykres dystrybuanty empirycznej odpowiadający obserwacji x, zamieszczony na rysunku 7, został uzyskany z następującego polecenia.

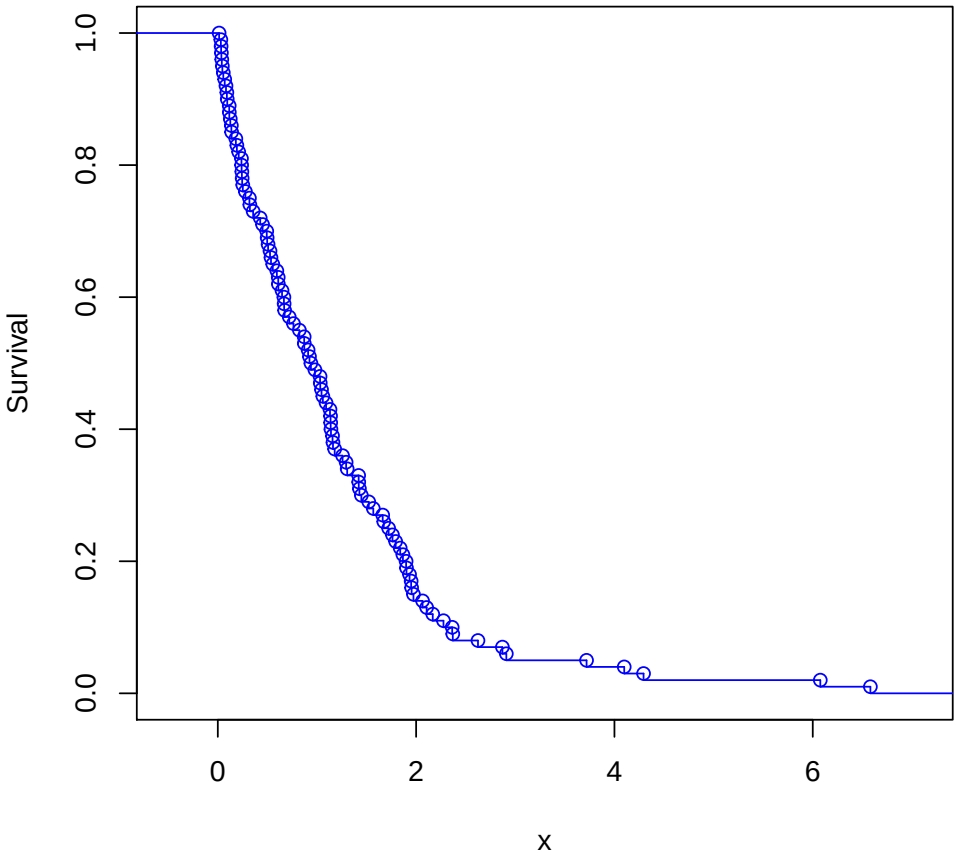
```
> plot(ecdf(x), main="",col="blue")
```



Rysunek 7: Wykres dystrybuanty epirycznej odpowiadający danym x

Wykres estymatora funkcji przeżycia (zobacz rysunek 8), odpowiadający wektorowi x, możemy uzyskać natępująco.

```
> x=sort(x)
> n=length(x)
> z=seq(from=0,to=1,by=1/n)
> y=1-z
> survival=stepfun(x,y,right=TRUE)
> plot(survival,ylab="Survival",main='',col="blue")
```



Rysunek 8: Wykres estymatora funkcji przeżycia na podstawie wektora x

4 Zadania do sprawozdania 1, część 1

4.1 zabawa

1. Narysować 2 wykresy funkcji hazardu rozkładu $BS(\alpha, \beta)$ odpowiadające wybranym parametrom tego rozkładu, dla których w jednym przypadku funkcja hazardu jest rosnąca, a w drugim – ma kształt wannowy.

2. Napisać program do generowania zmiennych z rozkładu $BS(\alpha, \beta)$.

3. Wygenerować $n = 10$ liczb z rozkładu $BS(\alpha, \beta)$ o wybranych parametrach i narysować na jednym wykresie teoretyczną dystrybuantę wybranego rozkładu i dystrybuantę empiryczną odpowiadającą wygenerowanym danym.

4. Wygenerować $n = 100$ liczb z rozkładu $BS(\alpha, \beta)$ o wybranych parametrach, wyznaczyć wartości podstawowych statystyk opisowych takich jak: średnia, mediana, odchylenie standardowe, kwartyl dolny, kwartyl górny, rozstęp, minimum i maksimum. Zilustrować wygenerowane dane na wybranych wykresach.

Literatura

[1] Kabacoff Robert, 2011. *R in Action*, Manning, Shelter Island.

[2] Paradis Emmanuel, 2005. *R for Beginners*, https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf.

5 Zadanie do sprawozdania - Część 1

5.1 Zadanie 1

Zadanie jest narysowanie 2 wykresóW funkcji hazardu rozkładu $BS(\alpha, \beta)$, gdzie $\alpha > 0, \beta > 0$ są parametrami kształtu. Biorąc $\beta >= 1$ dostaniemy rosnącą funkcje hazardu, poniżej znajduje się przykład(zobacz rysunek **??**), natomiast dla $0 < \beta < 1$ będzie miała kształt wannowy.

```
> x <- rchen(10, 1, 2)
> y <- rchen(10, 1, 0.6)
> t<-c(1,2,3,4,5,6)
>
>
```