

1.1 Rozkłady zmiennych losowych, czyli krótkie przypomnienie rachunku prawdopodobieństwa

UWAGA!!!

Nie należy bać się terminu zmienna losowa (a tym samym kończyć lektury tego rozdziału w tym miejscu) z dwóch powodów:

- biegłość w posługiwaniu się tym terminem jest wymagana do zaliczenia,
- brzmi nieciekawie, ale zaraz będzie wszystko jasne.

Na potrzeby kolejnych rozważań uznajmy, że

Definicja 1 *Zmienna losowa to funkcja przekształcająca wynik eksperymentu losowego na liczbę rzeczywistą.*

Pojęcie zmiennej losowej jest bardzo użyteczne, pozwala na abstrahowanie od postaci przestrzeni zdarzeń a operowanie wyłącznie na liczbach.

Wyobraźmy sobie rzut kostką sześciocinną. Jest to przykład eksperymentu losowego, czyli eksperymentu, dla którego wiemy jakie sytuacje mogą się wydarzyć, ale nie wiemy która się wydarzy. Możliwe wyniki eksperymentu są najróżniejsze, np: kostka będzie turlała się przez 2 minuty i wypadnie sześć oczek, kostka spadnie ze stołu i wypadnie jedno oczko itp.

Zmienna losowa to funkcja przedstawiająca wynik eksperymentu w postaci liczby rzeczywistej. Rozważając eksperyment rzutu kostką nie jest dla nas istotne co działo się z kostką, istotne jest jedynie ile oczek było na górnej ścianie kostki po jej zatrzymaniu (innymi słowy ile oczek wypadło). Dlatego naturalnym wyborem zmiennej losowej jest funkcja określająca liczbę oczek które wypadły w wyniku rzutu kostką.

Z uwagi na dalsze rozważania interesującą nas charakterystyką zmiennej losowej jest jej rozkład. Rozkład zmiennej losowej określa z jakim prawdopodobieństwem zmienna losowa przyjmuje poszczególne wartości.

Rozkład zmiennej losowej może być

- dyskretny (jeżeli zmienna losowa może przyjmować skończenie wiele lub przeliczalnie wiele wartości),
- ciągły,
- ani dyskretny ani ciągły (do tej klasy należą np. mieszaniny rozkładów dyskretnych u ciągłych. Nie będziemy w tym rozdziale rozważali rozkładów z tej klasy).

Formalnie zmienną losową definiuje się jako funkcję mierzalną z przestrzeni probabilistycznej do przestrzeni liczb rzeczywistych. Jednak na potrzeby tych zajęć do zrozumienia tematu nie potrzebujemy takich formalizmów. Jeżeli ktoś chciałby zgłębić swoją wiedzę na ten temat, to powinien rozważyć lekturę dobrej książki do rachunku prawdopodobieństwa.

Pewnie zastanawiacie się jaka będzie wartość tej zmiennej losowej jeżeli turlającą się kostkę zje kot? Dla uproszczenia i bez straty ogólności wykluczamy to zdarzenie z rozważań.

Rozkłady ciągłe można opisać funkcją nazywaną gęstością. Rozkłady dyskretne można opisać podając prawdopodobieństwo wystąpienia każdego z możliwych zdarzeń.

Dla każdego rozkładu można określić następujące wartości.

- dystrybuanta w punkcie x , to prawdopodobieństwo, że zmienna losowa przyjmie wartość mniejszą lub równą x ,
- prawdopodobieństwo w punkcie x (dla rozkładów dyskretnych), to prawdopodobieństwo, że zmienna losowa przyjmie wartość x ,
- gęstość w punkcie x (dla rozkładów ciągłych), to pochodna dystrybuanty,
- kwantyl w punkcie q , to wartość x , dla której prawdopodobieństwo, że zmienna losowa przyjmie wartość mniejszą lub równą x jest równe q (kwantyl to funkcja odwrotna dla dystrybuanty).

1.1.1 Normalny przykład

Jednym z najbardziej popularnych rozkładów jest rozkład normalny, nazywany też rozkładem gaussowskim. Na Rys 2. przedstawiona jest gęstość i dystrybuanta standardowego rozkładu normalnego .

czy jeżeli coś jest standardowe to jest też normalne?

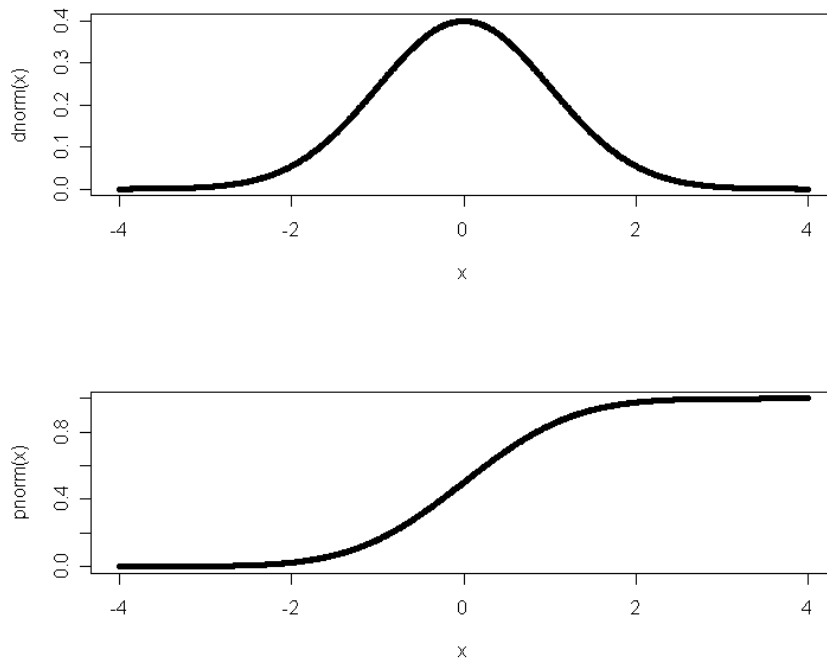
Kod generujący wykres Rys 2.

```
par(mfrow = c(2, 1))
x = seq(-4,4,by=0.01)
plot(x,dnorm(x),type="l",lwd=5)
plot(x,pnorm(x),type="l",lwd=5)
```

Rozkład normalny opisywany jest dwoma parametrami, średnią i wariancją. Standardowy rozkład normalny ma średnią równą 0 i wariancję równą 1. Przyglądając się Rysunkowi 2 można zauważyć, że prawdopodobieństwo, że zmienna losowa o rozkładzie normalnym przyjmie wartość mniejszą od -3 lub większą od 3 jest bardzo małe. Jeżeli pamiętamy czym jest dystrybuanta, oraz wiemy, że wartości dystrybuanty w R wyznaczyć można funkcją `pnorm()`, to z łatwością wyznaczymy to prawdopodobieństwo

```
> pnorm(-3) + (1 - pnorm(3))
[1] 0.002699796
```

Funkcje generujące liczby losowe z zadanego rozkładu mają prefix `r`. Lista



Rysunek 1.1: Gęstość i dystrybuanta standardowego rozkładu normalnego na odcinku $[-4, 4]$.

generatorów dla różnych rozkładów przedstawiona jest w następujących dwóch tabelach. Aby wylosować 10 liczb z rozkładu normalnego ze średnią równą 2 i wariancją równą 1 należy wpisać

nie gwarantuje, że
otrzymacie ten sam
wynik

```
> rnorm(10,2,1)
[1] 0.6768987 1.2803642 2.2650092 0.6985579 2.3181149
[6] 1.8539709 3.0213374 1.6717404 2.9371182 1.5042659
```

W poniższych tabelach wylistowano funkcje służące do wyznaczania wartości dystrybuant, gęstości oraz kwantyli dla poszczególnych rozkładów.

1.1.2 Rozkłady ciągłe

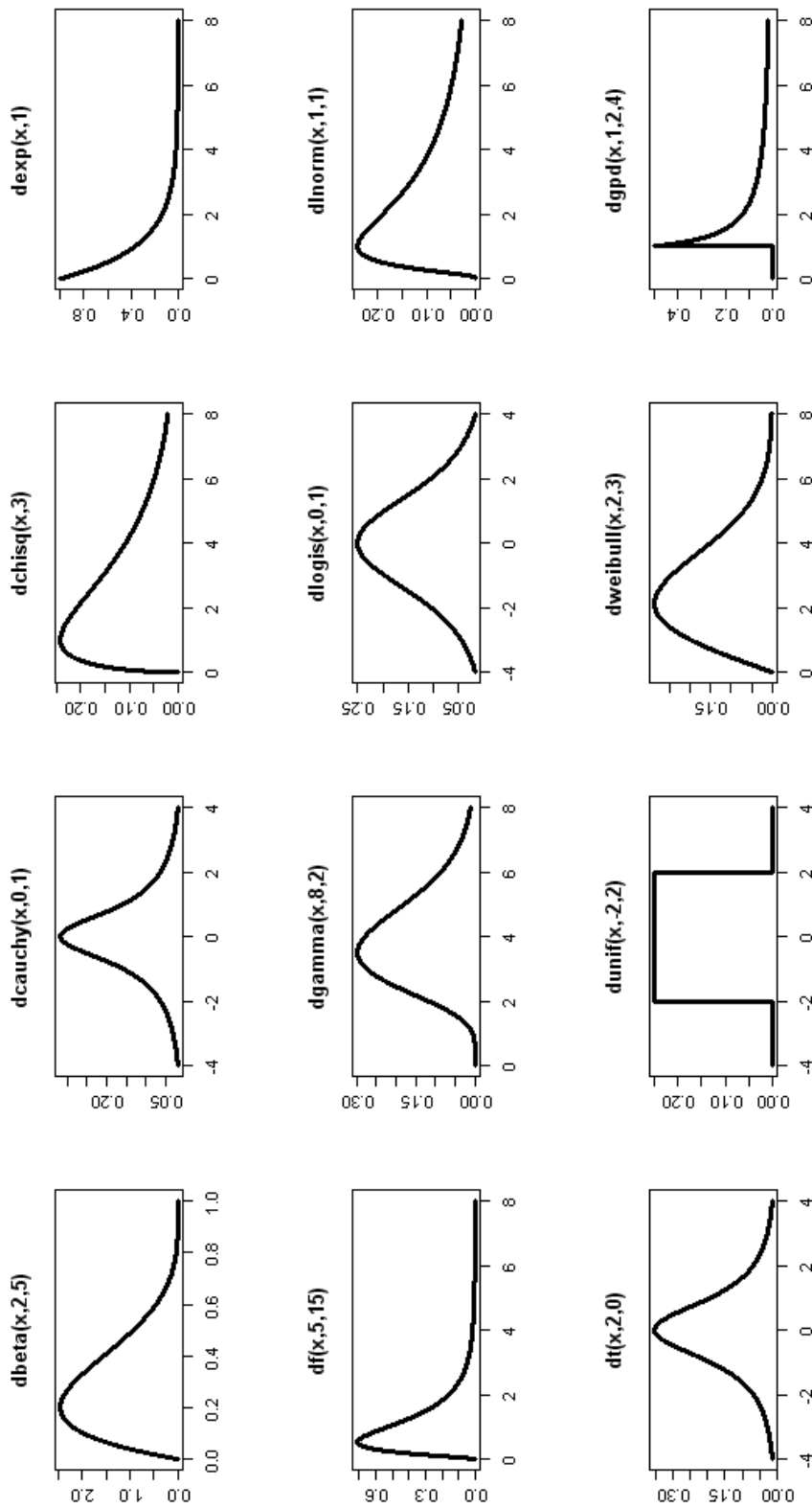
TODO: Opisać te rozkłady!

nazwa	dystybuanta	kwantyl	gęstość	generator	parametry	pakiet
Beta	pbeta	qbeta	dbeta	rbeta	shape1, shape2	stats
Cauchy	pcauchy	qcauchy	dcauchy	rcauchy	location, scale	stats
Chi-Square	pchisq	qchisq	dchisq	rchisq	df, ncp	stats
Exponential	pexp	qexp	dexp	rexp	rate	stats
F	pf	qf	df	rf	df1, df2, ncp	stats
Gamma	pgamma	qgamma	dgamma	rgamma	shape, rate	stats
Logistic	plogis	qlogis	dlogis	rlogis	location, scale	stats
LogNormal	plnorm	qlnorm	dlnorm	rlnorm	meanlog, sdlog	stats
Normal	pnorm	qnorm	dnorm	rnorm	mean, sd	stats
Student t	pt	qt	dt	rt	df, ncp	stats
Studentized Range	ptukey	qtukey			nmeans, df, nranges	stats
Uniform	punif	qunif	dunif	runif	min, max	stats
Weibull	pweibull	qweibull	dweibull	rweibull	shape, scale	stats
Pareto	pgpd	qgpd	dgpd	rgpd	loc, scale, shape	evd, evir

1.1.3 Rozkłady dyskretne

TODO: Opisać te rozkłady!

nazwa	dystrybuanta	kwantyl	gęstość	generator	parametry	pakiet
Binomial	pbinom	qbinom	dbinom	rbinom	size, prob	stats
Geometric	pgeom	qgeom	dgeom	rgeom	prob	stats
Hypergeometric	phyper	qhyper	dhyper	rhyper	m,n,k	stats
Negative Binomial	pnbinom	qnbinom	dnbinom	rnbinom	size, prob	stats
Poisson	ppois	qpois	dpois	rpois	lambda	stats
Wilcoxon Rank Sum Statistic	pwilcox	qwilcox	dwilcox	rwilcox	n,m	stats
Wilcoxon Signed Rank Statistic	psignrank	qsignrank	dsignrank	rsignrank	n	stats



Rysunek 1.2: Funkcje gęstości dla przykładowych parametrów różnych rozkładów ciągłych.

1.2 Zadania

1. 30 tys studentów politechniki zdaje test ze znajomości budowy kosy. Test ma 100 pytań. Na każde z pytań osoba nieznająca odpowiedzi może odpowiedzieć poprawnie z prawdopodobieństwem 0.25. Zakładając, że żaden student nie zna odpowiedzi na żadne pytanie, oceń ilu średnio studentów zda test, jeżeli do zdania wystarczy otrzymać 30 poprawnych odpowiedzi (oczywiście należy napisać odpowiedni program w R).
2. Korzystając z funkcji wyznaczającej kwantyle wylicz jak powinien zostać dobrana minimalna liczba poprawnych odpowiedzi niezbędnych do zdania, by test zdało nie więcej niż 5%.
3. Wyznacz średnią z 100 tys obserwacji pochodzących z rozkładu normalnego ze średnią 2 i wariancją 1, cauchego z parametrem położenia 2 i skali 1 i rozkładu log-normalnego ze średnią 2 i wariancją 1. Powtórz trzykrotnie losowanie oraz wyznaczanie średniej i porównaj wyniki. Która ze średnich zachowuje się dziwnie? Dlaczego?
4. Znajdź taką wartość x by prawdopodobieństwo że zmienna losowa o rozkładzie standardowym normalnym znajduje się w przedziale $[-x, x]$ było równe 0.9. Znajdź taką wartość również dla rozkładów t-Studenta i logistycznego.