Statystyka w analizie i planowaniu eksperymentu

Paweł Błażej

9 marca 2016

Zmienna losowa - wstęp

Przeprowadzane w praktyce badania i eksperymenty mają bardzo różnorodny charakter, niemniej jednak wiążą się one z rejestracją jakiś sygnałów (danych). Mogą to być na przykład:

1. odczyty na skali;

Zmienna losowa - wstęp

Przeprowadzane w praktyce badania i eksperymenty mają bardzo różnorodny charakter, niemniej jednak wiążą się one z rejestracją jakiś sygnałów (danych). Mogą to być na przykład:

1. odczyty na skali;
2. końcowe parametry jakiegoś procesu technologicznego;

Zmienna losowa - wstęp

Przeprowadzane w praktyce badania i eksperymenty mają bardzo różnorodny charakter, niemniej jednak wiążą się one z rejestracją jakiś sygnałów (danych). Mogą to być na przykład:

1. odczyty na skali;
2. końcowe parametry jakiegoś procesu technologicznego;
3. liczba osób w kolejce.

Klasyfikacja zmiennych Stevensa

1. Zmienne jakościowe (nazywane również kategorycznymi, czynnikowymi), to zmienne przyjmujące określoną liczbę wartości (najczęściej nieliczbowych):
   - binarne, np. płeć (kobieta/mężczyzna);
   - nominalne, np. marka samochodu;
   - porządkowe, np. wykształcenie (podstawowe, średnie, wyższe).

Klasyfikacja zmiennych Stevensa

1. Zmienne jakościowe (nazywane również kategorycznymi, czynnikowymi), to zmienne przyjmujące określoną liczbę wartości (najczęściej nieliczbowych):
   - binarne, np. płeć (kobieta/mężczyzna);
   - nominalne, np. marka samochodu;
   - porządkowe, np. wykształcenie (podstawowe, średnie, wyższe).
2. Zmienne ilościowe, opisują ilość. Wyróżnia się skale:
   - licznikową (liczebność wystąpień pewnego zjawiska, opisywana przez liczby naturalne) np. liczba lat nauki;
   - przedziałową (interwałową) skala, w której zmienna może przyjmować dowolne wartości z określonego przedziału.

Zmienna losowa i jej dystrybuanta

Tej definicji nie trzeba uczyć się na pamięć!

Zmienną losową nazywamy funkcję X określoną na przestrzeni zdarzeń elementarnych Ω, o wartościach ze zbioru liczb rzeczywistych $\mathbf{R}$.

Uwaga

Zmienne losowe oznaczamy dużymi literami X, Y, Z a ich konkretne wartości małymi literami x, y, z.

Dystrybuanta zmiennej losowej i jej własności

Dystrybuanta zmiennej losowej X

Funkcję F_X określoną na zbiorze $\mathbf{R} = (-\infty, +\infty)$ liczb rzeczywistych wzorem:

$$F_X(x) = P(X \leqslant x), \ x \in \mathbf{R}$$

nazywamy dystrybuantą zmiennej losowej X.

Jeżeli nie ma wątpliwości z jaką zmienną losową mamy do czynienia wtedy dystrybuantę oznaczamy przez F.

Przykład cd.

Załóżmy, że prawdopodobieństwo wylosowania wadliwego towaru wynosi $0 < p < 1$. Wówczas dystrybuanta zmiennej losowej X przyjmuje postać

$$F(x) = \begin{cases} 0, & \text{dla} \quad x < 0 \\ 1 - p & \text{dla} \quad 0 \leqslant x < 1 \\ 1 & \text{dla} \quad x \geqslant 1 \end{cases}$$

Własności dystrybuanty

1. $0 \leqslant F(x) \leqslant 1$ dla każdego $x \in \mathbf{R}$;

Własności dystrybuanty

1. $0 \leqslant F(x) \leqslant 1$ dla każdego $x \in \mathbf{R}$;
2. $\lim_{x \to -\infty} F(x) = 0$ oraz $\lim_{x \to +\infty} F(x) = 1$;

Własności dystrybuanty

1. $0 \leqslant F(x) \leqslant 1$ dla każdego $x \in \mathbf{R}$;
2. $\lim_{x \to -\infty} F(x) = 0$ oraz $\lim_{x \to +\infty} F(x) = 1$;
3. jest funkcją niemalejącą;

Własności dystrybuanty

1. $0 \leqslant F(x) \leqslant 1$ dla każdego $x \in \mathbf{R}$;
2. $\lim_{x \to -\infty} F(x) = 0$ oraz $\lim_{x \to +\infty} F(x) = 1$;
3. jest funkcją niemalejącą;
4. $P(a < X \leqslant b) = F(b) - F(a)$

Dwa ważne typy zmiennych losowych

Zmienna losowa typu dyskretnego

Mówimy, że zmienna losowa X jest typu skokowego (dyskretnego) jeżeli istnieje skończony albo przeliczalny zbiór $W_x = \{x_1, x_2, \ldots, \ldots\}$ jej wartości taki, że

$$P(X = x_i) = p_i, \ \ i \in \mathbf{N},$$

$$\sum_{i=1} p_i = 1,$$

gdzie górna granica sumowania wynosi n albo ∞

Dwa ważne typy zmiennych losowych

Funkcja prawdopodobieństwa

Funkcję p przyjmującą wartości $p(x_i) = P(X = x_i)$ oznaczaną często przez p_i nazywamy funkcją prawdopodobieństwa zmiennej losowej X.

Dystrybuanta zmiennej losowej typu dyskretnego

Gdy dane jest funkcja prawdopodobieństwa zmiennej losowej X, to prawdopodobieństwo przyjęcia przez tą zmienną wartości ze zbioru A jest określone równością:

$$P(X \in A) = \sum_{x_i \in A} p_i$$

W szczególności dla dowolnego przedziału (a, b) zachodzi

$$P(-\infty < X < b) = \sum_{-\infty < x_i < b} p_i$$

Wybrane charakterystyki rozkładów - przypadek dyskretny

1. wartość średnia (parametr położenia)

$$\mu = \sum_{i=1}^{k} x_i p_i;$$

2. wariancja (parametr skali)

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 p_i.$$

Ważne rozkłady typu dyskretnego

1. Rozkład Bernoulliego (rzut monetą);

Ważne rozkłady typu dyskretnego

1. Rozkład Bernoulliego (rzut monetą);
2. Rozkład dwumianowy (rzut wieloma monetami);

Ważne rozkłady typu dyskretnego

1. Rozkład Bernoulliego (rzut monetą);
2. Rozkład dwumianowy (rzut wieloma monetami);
3. Rozkład Poissona (liczba sygnałów);

Ważne rozkłady typu dyskretnego

1. Rozkład Bernoulliego (rzut monetą);
2. Rozkład dwumianowy (rzut wieloma monetami);
3. Rozkład Poissona (liczba sygnałów);
4. Rozkład hipergeometryczny (ryby w stawie).

Przykłady rozkładów dyskretnych

Rozkład Bernoulliego

Zmienna losowa X ma rozkład Bernoulliego, jeżeli jej funkcja prawdopodobieństwa jest postaci

x_i	0	1
p_i	q	p

$$\mu = p$$

$$\sigma^2 = pq$$

Rozkład dwumianowy

Zmienna losowa X ma rozkład dwumianowy z parametrami (n, p), $n \in \mathbf{N}$, $0 < p < 1$, jeżeli jej funkcja prawdopodobieństwa jest postaci

$$P(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mu = np$$

$$\sigma^2 = npq$$

Zadanie

Rzucamy 6 razy symetryczną monetą oblicz prawdopodobieństwo wyrzucenia conajmniej jednego „orła”

Rozkład hipergeometryczny

Zmienna losowa X ma rozkład hipergeometryczny z parametrami (N, M, n), gdzie N, M, n to liczby naturalne oraz $M, n \leqslant N$, jeżeli jej funkcja prawdopodobieństwa jest postaci

$$P(k; N, M, n) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

$$\mu = np$$

$$\sigma^2 = npq\frac{N-n}{N-1}$$

Zadanie

W stawie jest $N = 100$ ryb odławiamy $M = 40$ z nich znakujemy je i z powrotem wrzucamy do stawu. Następnie łowimy $n = 20$ sztuk. Jakie jest prawdopodobieństwo, że wsród nich będzie dokładnie k oznakowanych?

Rozkład Poissona

Zmienna losowa X ma rozkład Poissona z parametrem λ, gdzie $\lambda > 0$, jeżeli jej funkcja prawdopodobieństwa jest postaci

$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\mu = \lambda$$

$$\sigma^2 = \lambda$$

Rozkład Poissona jest związany z sytuacją zliczania zdarzeń losowych określonego typu w pewnym odcinku czasu. Może to być na przykład zliczenie ilości kolejnych klientów pojawiających się w kasie w banku, ilość samochodów przejeżdżających punkt kontrolny.

Twierdzenie Poissona

Jeżeli $n \to \infty$, $p_n \to 0$, $np_n \to \lambda$, to

$$\binom{n}{k} p_n^k (1 - p_n)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}$$

Twierdzenie Poissona zastosowania

Prawdopodobieństwo trafienia „szóstki ”w Toto-Lotku jest równe $1/\binom{49}{6} = 1/139883816$. Ilu szóstek należy się spodziewać w każdym tygodniu, jeżeli grający wypełniają kupony niezależnie od siebie i całkowicie losowo, kuponów jest $n = 10^7$?

Zmienne losowe typu ciągłego

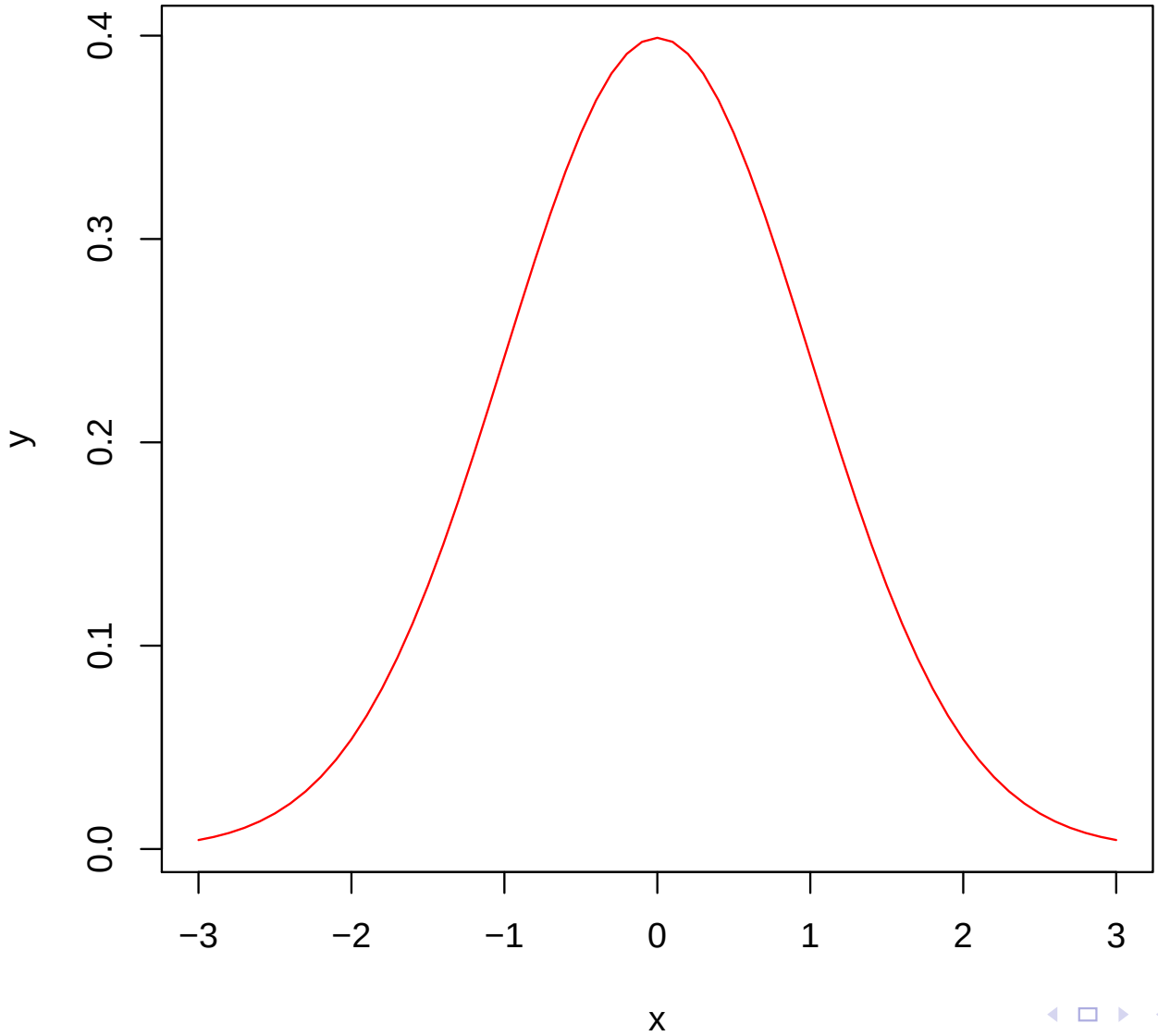
Zmienne losowe typu ciągłego mogą przyjmować nieskończenie wiele wartości (np. temperatura powietrza, waga ciała, prędkość samochodu w zadanej chwili czasu)

Zmienna losowa typu ciągłego

Funkcja gęstości zmiennej losowej X typu ciągłego, to nieujemna funkcja f, taka, że dla każdego przedziału (x_1, x_2)

$$P(\{\omega : \ x_1 \leqslant X \leqslant x_2\}) = \int_{x_1}^{x_2} f(x)dx.$$

Gęstość zmiennej losowej typu ciągłego



Własności funkcji gęstości

1. $f(x) \geqslant 0$ dla każdego x;
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Wybrane rozkłady typu ciągłego

1. Rozkład normalny;
2. Rozkład χ^2;
3. Rozkład Studenta;
4. Rozkład F.

Przykłady rozkładów ciągłych

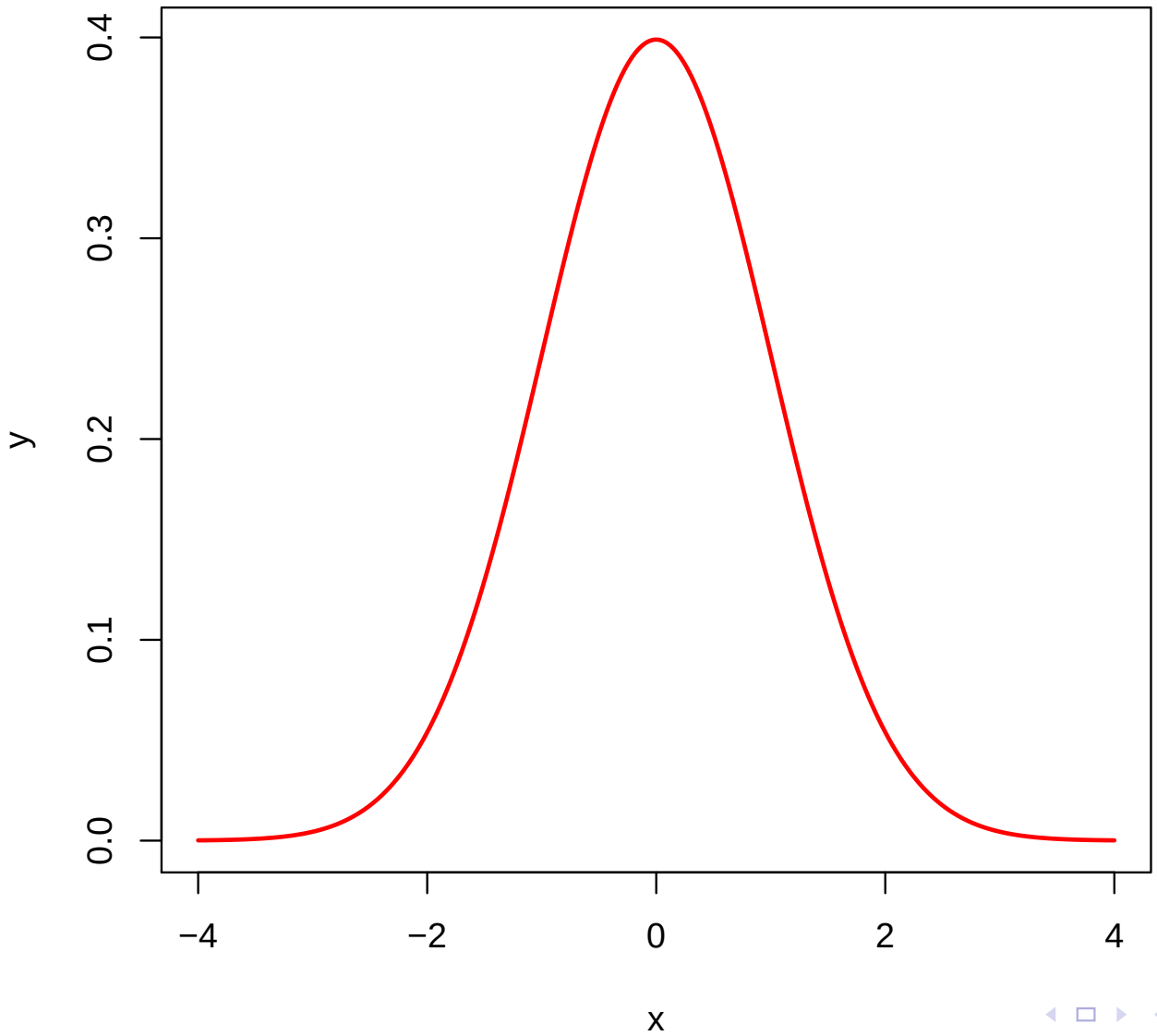
Rozkład normalny

Niech X będzie ciągłą zmienną losową o gęstości prawdopodobieństwa

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ x \in \mathbf{R},$$

gdzie $\mu \in \mathbf{R}$ i $\sigma > 0$ są danymi parametrami. Mówimy wtedy, że X ma *rozkład normalny* co oznaczamy $N(\mu, \sigma)$.

Funkcja gęstości rozkładu normalnego $N(\mu, \sigma)$.



Przykłady rozkładów ciągłych

Rozkład t-Studenta

Zmienna losowa X ma rozkład t-Studenta o n stopniach swobody, jeżeli jej gęstość f wyraża się wzorem

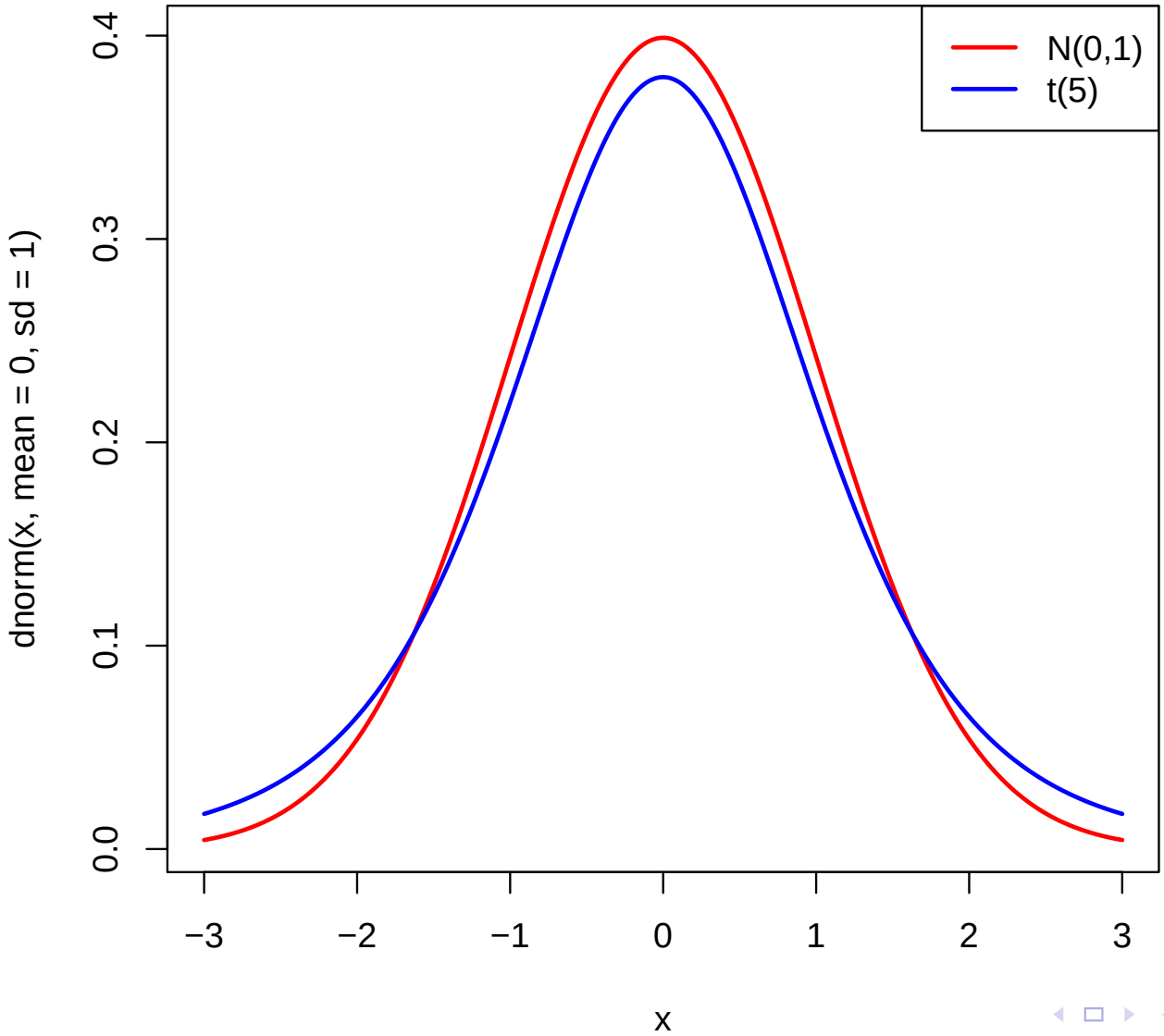
$$f(x) = \frac{\Gamma([n+1]/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+\frac{x^2}{n})^{-(n+1)/2}, x \in \mathbf{R}, n \in \mathbf{N}.$$

gdzie $\Gamma(n) = \int_0^{+\infty} x^{n-1}e^{-x}dx, \ n > 0$

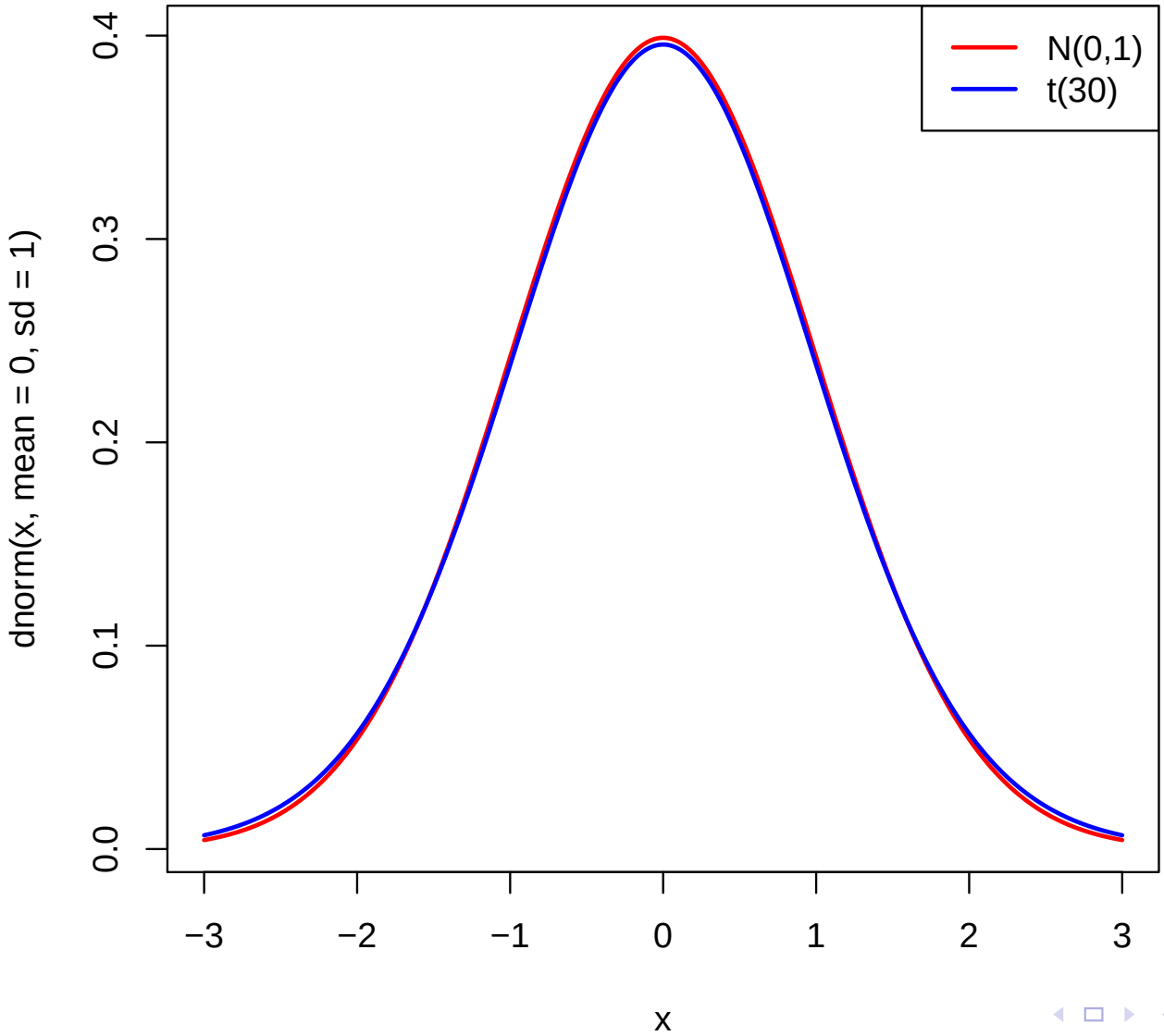
Uwaga

Gdy $n \to \infty$, gęstość f „dąży" do gęstości rozkładu normalnego $N(0,1)$, co wykorzystuje się w praktyce dla $n \geqslant 30$ do przybliżania rozkładu Studenta rozkładem normalnym.

Wykresy gęstości rozkładu Studenta i rozkładu normalnego



Wykresy gęstości rozkładu Studenta i rozkładu normalnego



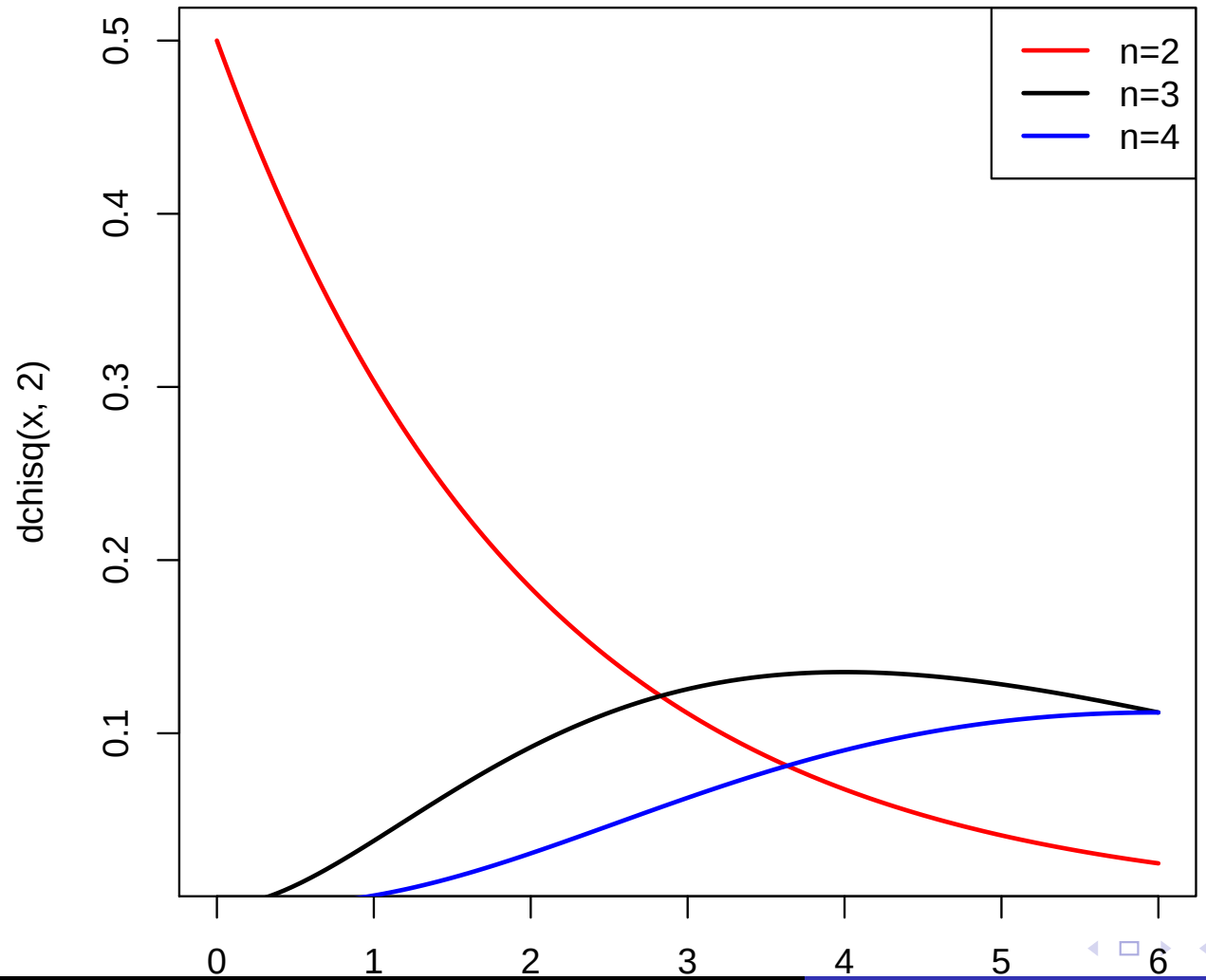
Przykłady rozkładów ciągłych

Rozkład χ^2

Zmienna losowa X ma rozkład χ^2 o n stopniach swobody, jeżeli jej gęstość f wyraża się wzorem

$$f(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} & \text{gdy } x > 0 \\ 0 & x \leqslant 0 \end{cases}$$

Wykresy gęstości rozkładu χ^2 o $n = 2, 6, 8$ stopniach swobody



Własności rozkładu normalnego

Rozkład normalny jest rozkładem najczęściej wykorzystywanym do modelowania zmienności pewnej cechy w zadanej populacji. Parametrami tego rozkładu są średnia μ (parametr położenia) oraz σ (parametr skali).

Własności rozkładu normalnego

Zapis

$$X \sim \mathcal{N}(\mu, \sigma).$$

Standardowy rozkład normalny

$$X \sim \mathcal{N}(0, 1), \ \textit{gdzie} \ \mu = 0, \ \sigma = 1$$

Przekształcenia zmiennej losowej o rozkładzie normalnym

Niech

$$X \sim \mathcal{N}(\mu, \sigma),$$

wtedy

$$Y \sim \mathcal{N}(0, 1), \text{ gdzie } Y = \frac{(X - \mu)}{\sigma}.$$

Przekształcenia zmiennej losowej o rozkładzie normalnym

Niech

$$X \sim \mathcal{N}(0,1),$$

wtedy

$$Y \sim \mathcal{N}(\mu,\sigma) \text{ gdzie } Y = \sigma \cdot X + \mu$$

Gęstość rozkładu normalnego

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Podstawowe własności rozkładu normalnego

1. $E[X] = \mu$;

Podstawowe własności rozkładu normalnego

1. $E[X] = \mu$;
2. $Var[X] = \sigma^2$;

Podstawowe własności rozkładu normalnego

1. $E[X] = \mu$;
2. $Var[X] = \sigma^2$;
3. Jeżeli $X_1, X_2, \ldots, X_n$ są to niezależne zmienne losowe pochodzące z rozkładu normalnego $N(\mu, \sigma)$. to $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$;

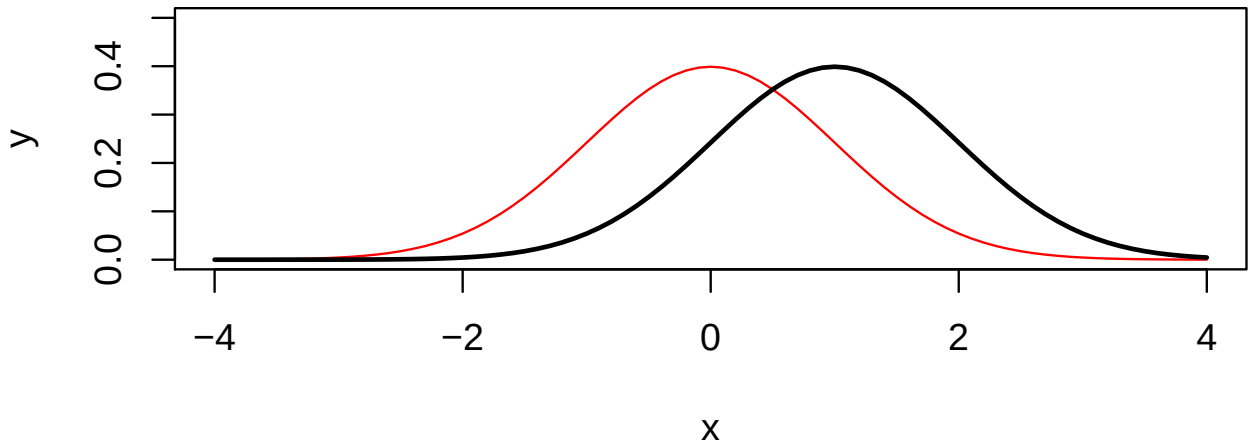
Podstawowe własności rozkładu normalnego

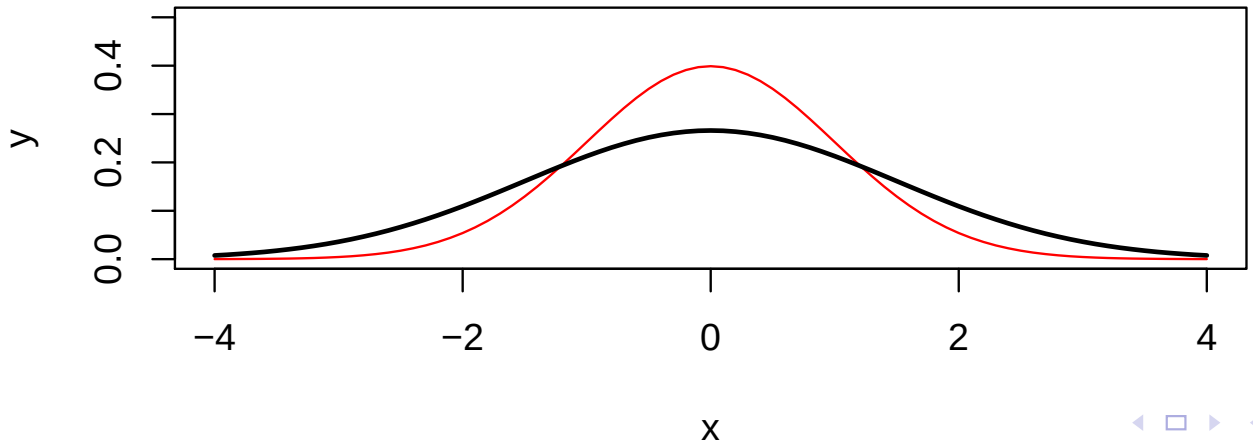
1. $E[X] = \mu$;
2. $Var[X] = \sigma^2$;
3. Jeżeli $X_1, X_2, \ldots, X_n$ są to niezależne zmienne losowe pochodzące z rozkładu normalnego $N(\mu, \sigma)$. to $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$;
4. $Me = \mu$;

Podstawowe własności rozkładu normalnego

1. $E[X] = \mu$;
2. $Var[X] = \sigma^2$;
3. Jeżeli $X_1, X_2, \ldots, X_n$ są to niezależne zmienne losowe pochodzące z rozkładu normalnego $N(\mu, \sigma)$. to $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$;
4. $Me = \mu$;
5. $Mo = \mu$;

Różnice w charakterystykach położenia i skali



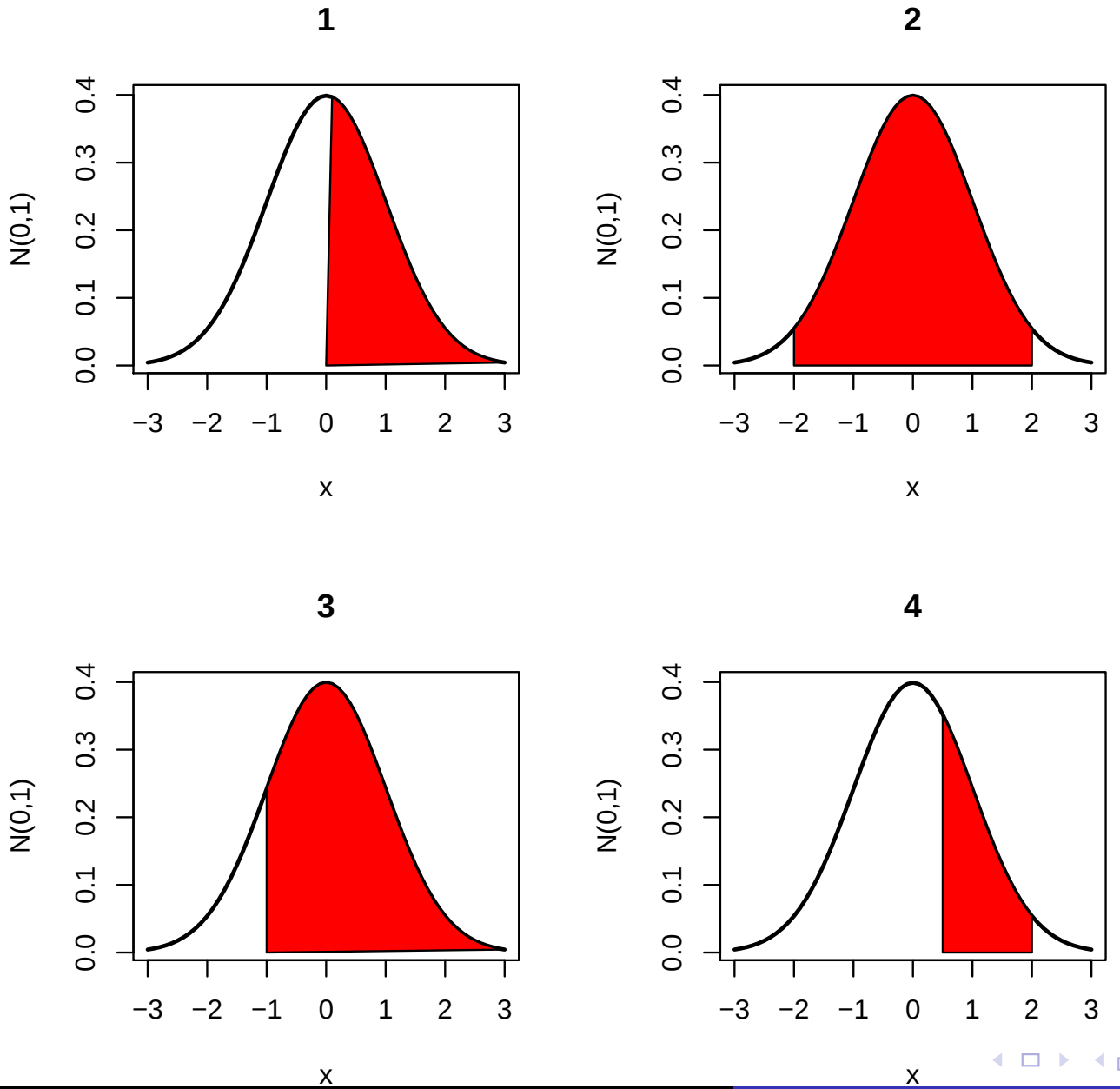


Praktyczne posługiwanie się rozkładem normalnym

Niech $X \sim \mathcal{N}(0, 1)$. Oblicz następujące prawdopodobieństwa:

1. $P(X > 0) = 0.5$;
2. $P(|X| < 2) = 0.954499736103642$;
3. $P(X > -1) = 0.841344746068543$;
4. $P(0.5 < X < 2) = 0.285787406777808$.

Praktyczne posługiwanie się rozkładem normalnym

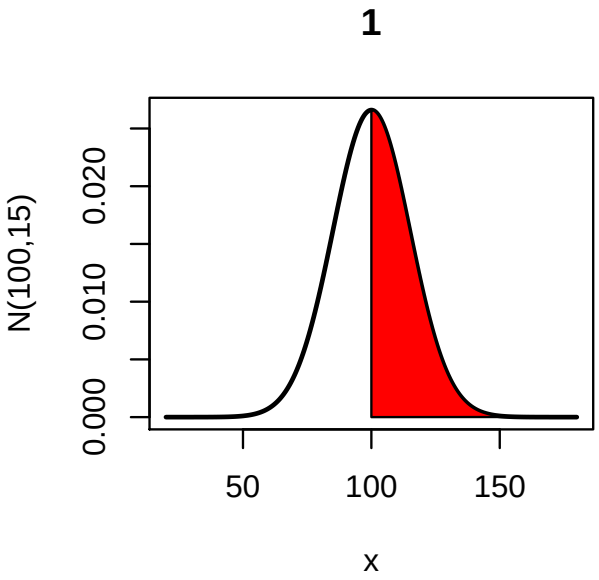


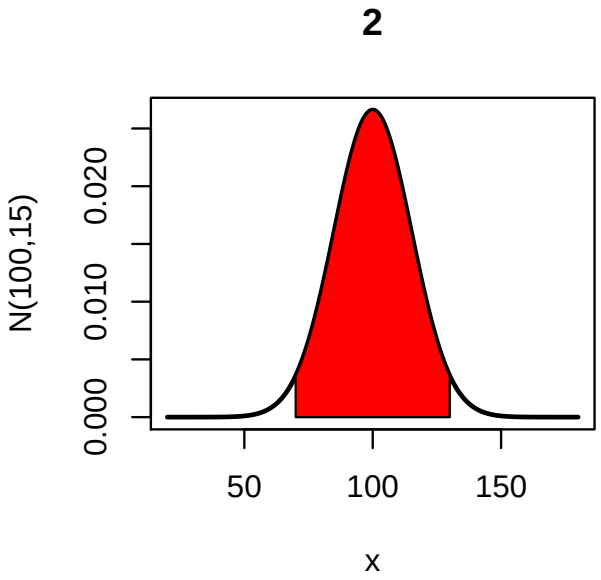
Praktyczne posługiwanie się rozkładem normalnym

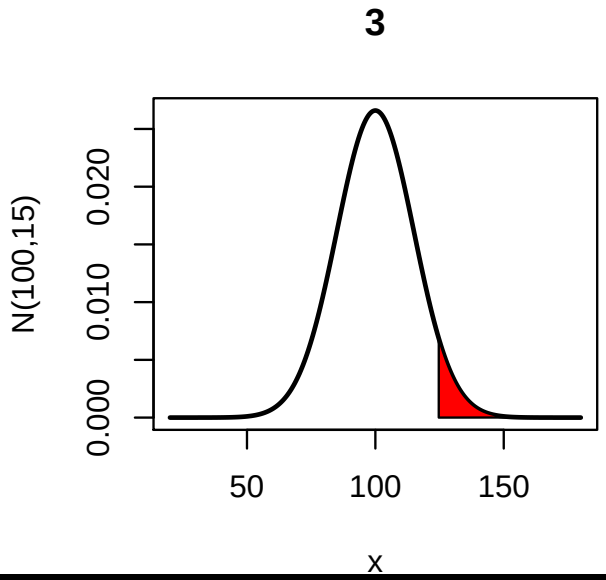
Przyjmuje się, że współczynnik IQ ma w populacji rozkład normalny o średniej $\mu = 100$ i odchyleniu standardowym $\sigma = 15$

1. Ile osób ma większe IQ niż 100?;
2. Ile osób ma IQ w przedziale $70 - 130$?;
3. Jaki przedział przyjąć aby określić odsetek 0.05 osób o największym IQ?

Praktyczne posługiwanie się rozkładem normalnym







Praktyczne posługiwanie się rozkładem normalnym - odpowiedzi

1. $P(X > 100) = 0.5$;
2. $P(70 < X < 130) = 0.954499736103642$;
3. $P(X > 124.75) = 0.05$.

Praktyczne posługiwanie się rozkładem normalnym

Załóżmy, że długość piór ogonowych pawia wynosi średnio 65 cm, z odchyleniem standardowym 5 cm. Oszacuj prawdopodobieństwo, że losowo wyjęte pióro ma długość:

1. mniejszą niż 54 cm;
2. większą niż 64 cm;
3. jeśli mieszkańcy Łobzowa zwykli nosić na czapkach pióra o długości od 70 do 75 cm, to jak często natrafiają na takie pióro?

Praktyczne posługiwanie się rozkładem normalnym - odpowiedzi

1. P(dł. mniejsza niż 54 cm) = 0.0139034475134986;
2. P(dł. większą niż 64 cm)=0.579259709439103;
3. P (70cm $<$ dł. $<$75cm)=0.135905121983278.

Reguła 3σ

Dla rozkładu normalnego ma zastosowanie tzw. „ prawo 3σ ” mówi ono, że:

1. ok. 68% wszystkich wartości zmiennej odbiega od średniej oczekiwanej nie bardziej niż o jedno odchylenie standardowe;

Reguła 3σ

Dla rozkładu normalnego ma zastosowanie tzw. „ prawo 3σ ” mówi ono, że:

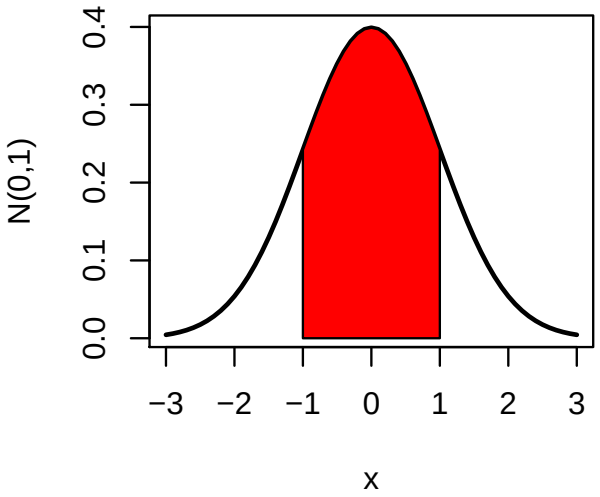
1. ok. 68% wszystkich wartości zmiennej odbiega od średniej oczekiwanej nie bardziej niż o jedno odchylenie standardowe;
2. ok. 95% wszystkich wartości nie bardziej niż o dwa odchylenia standardowe;

Reguła 3σ

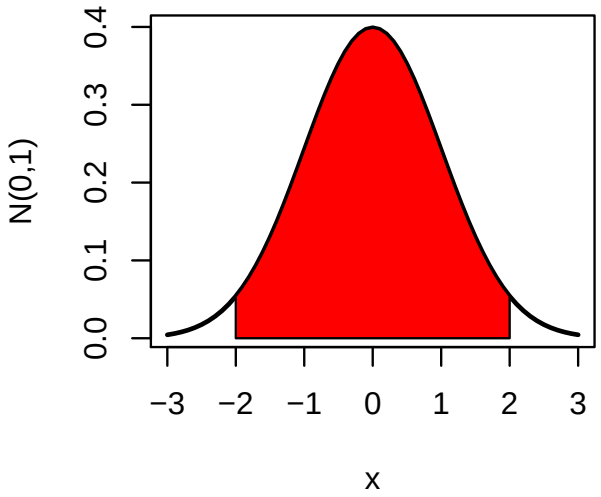
Dla rozkładu normalnego ma zastosowanie tzw. „ prawo 3σ ” mówi ono, że:

1. ok. 68% wszystkich wartości zmiennej odbiega od średniej oczekiwanej nie bardziej niż o jedno odchylenie standardowe;
2. ok. 95% wszystkich wartości nie bardziej niż o dwa odchylenia standardowe;
3. ok. 99.8% odbiega o nie więcej niż 3σ od wartości średniej.

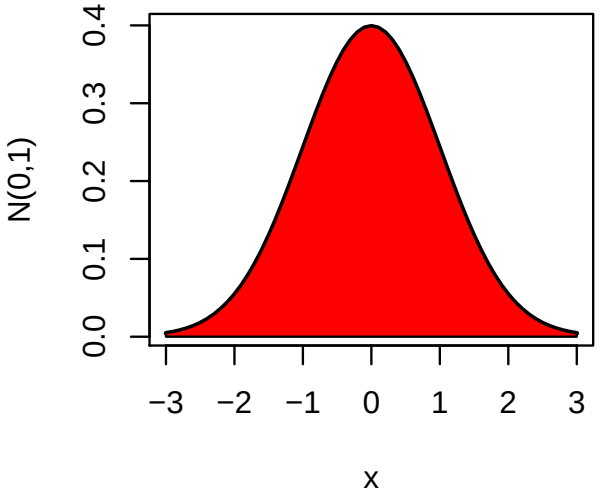
1


N(0,1)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x

2


N(0,1)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x

3


N(0,1)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x

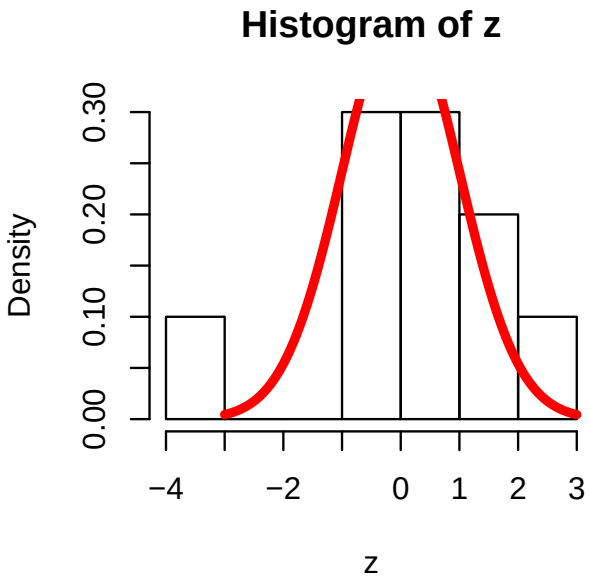
Centralne Twierdzenie Graniczne

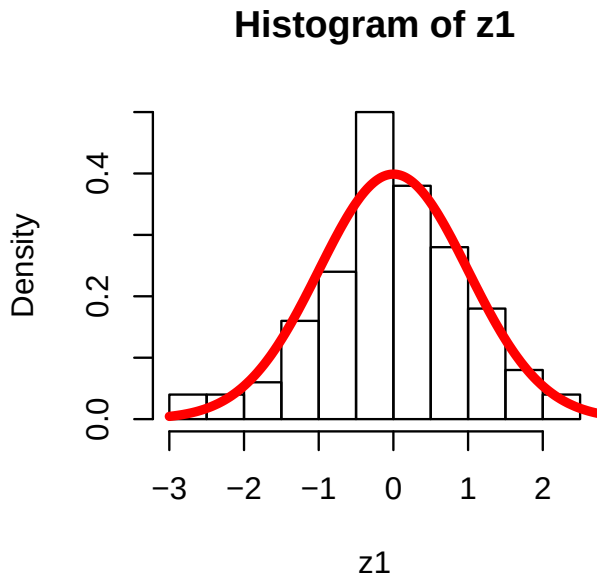
Rozkład normalny jest rozkładem granicznym dla standaryzowanych zmiennych losowych pochodzących z różnych rozkładów.

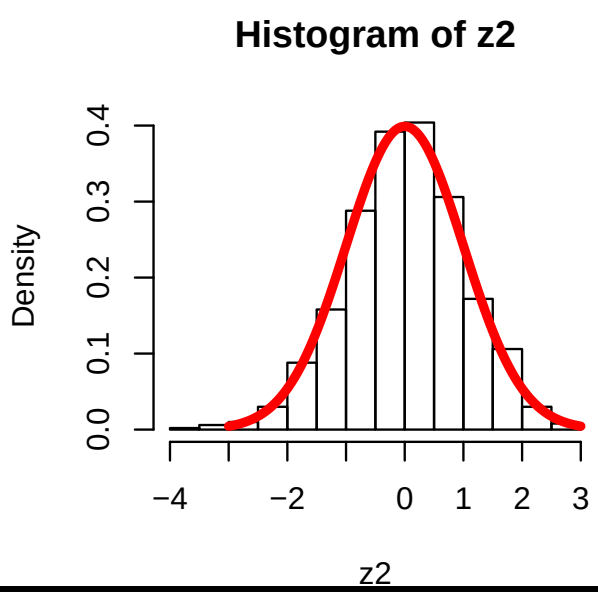
Centralne Twierdzenie Graniczne

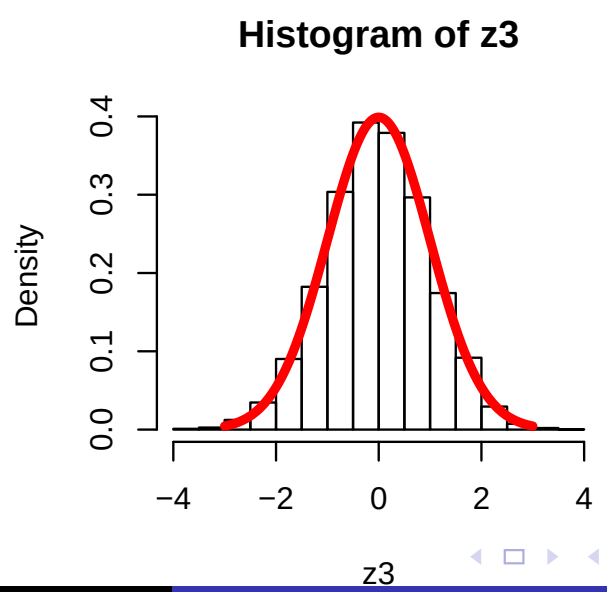
Średnia n niezależnych ustandaryzowanych $\sqrt{n}\frac{\bar{X}-E[X]}{\sqrt{Var[X]}}$ zmiennych losowych pochodzących z porządnych rozkładów zbiega do rozkładu normalnego $\mathcal{N}(0,1)$.

Co to oznacza w praktyce?









Praktyczne posługiwanie się rozkładem normalnym

Rzucamy 100 razy monetą jakie jest prawdopodobieństwo, że suma wyrzuconych orłów przekroczy 55?

Praktyczne posługiwanie się rozkładem normalnym

Rzucamy 20 razy symetryczną kostką oblicz prawdopodobieństwo tego, że suma wyrzuconych oczek jest liczbą pomiędzy $60, 80$?

darzenia, ze średni dzienny dojazd w ciągu 30 dni