ROZKŁADY PRAWDOPODOBIEŃSTWA ZMIENNEJ LlosowEJ

Statystyka biomedyczna

Artur Klepaczko

KLUCZOWE PYTANIE #1

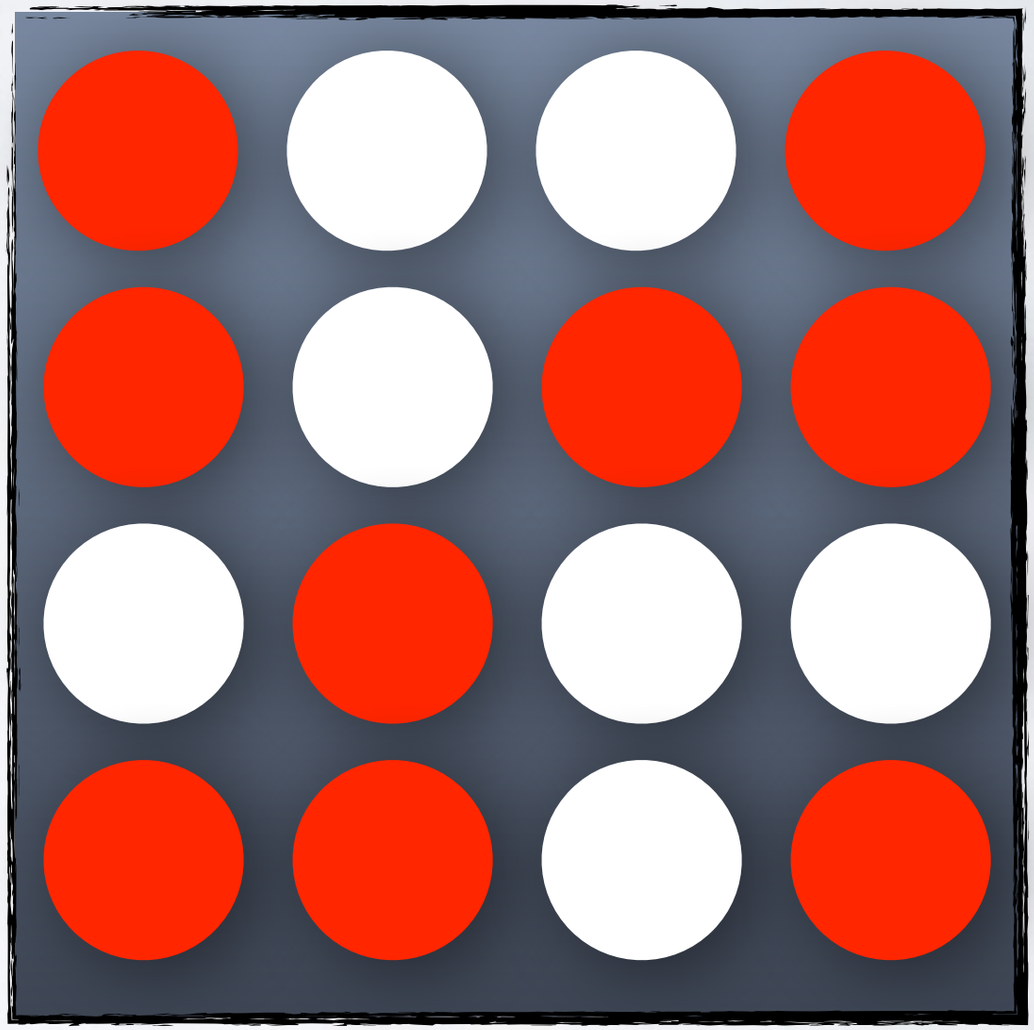
Co to jest zmienna losowa?

ZMIENNA LOSOWA

- Zmienna losowa
  - **kula, *X***

- Realizacja zmiennej losowej
  - **czerwona** kula, ***X* = *c***
  - **biała** kula, ***X* = *b***

- Prawdopodobieństwo
  - **Pr(*X*=*c*) = 9/16**
  - **Pr(*X*=*b*) = 7/16**



ZMIENNA LOSOWA

- Obserwowana cecha obiektów lub zjawisk o charakterze losowym (stochastycznym)
- Funkcja lub odwzorowanie obiektu w przestrzeń liczbową.
- Jeśli wartość zmiennej losowej należy do zbioru skończonego lub przeliczalnego, to mówimy o zmiennej dysretnej.
- Jeśli zmienna losowa przyjmuje dowolne wartości liczbowe, to jest to zmienna losowa ciągła.

KLUCZOWE PYTANIE #2

Co to jest rozkład prawdopodobieństwa zmiennej losowej?

ROZKŁAD PRAWDOPODOBIEŃSTWA

- Rozkład zmiennej losowej **dyskretnej** jest to **zbiór wartości prawdopodobieństwa** dla określonych realizacji tej zmiennej
- Ponieważ zmienna losowa ciągła przyjmuje dowolne wartości z określonego przedziału, prawdopodobieństwo jej realizacji w danym punkcie zawsze wynosi 0. Dlatego rozkład zmiennej losowej **ciągłej** opisujemy za pomocą **funckji gęstości prawdopodobieństwa**.

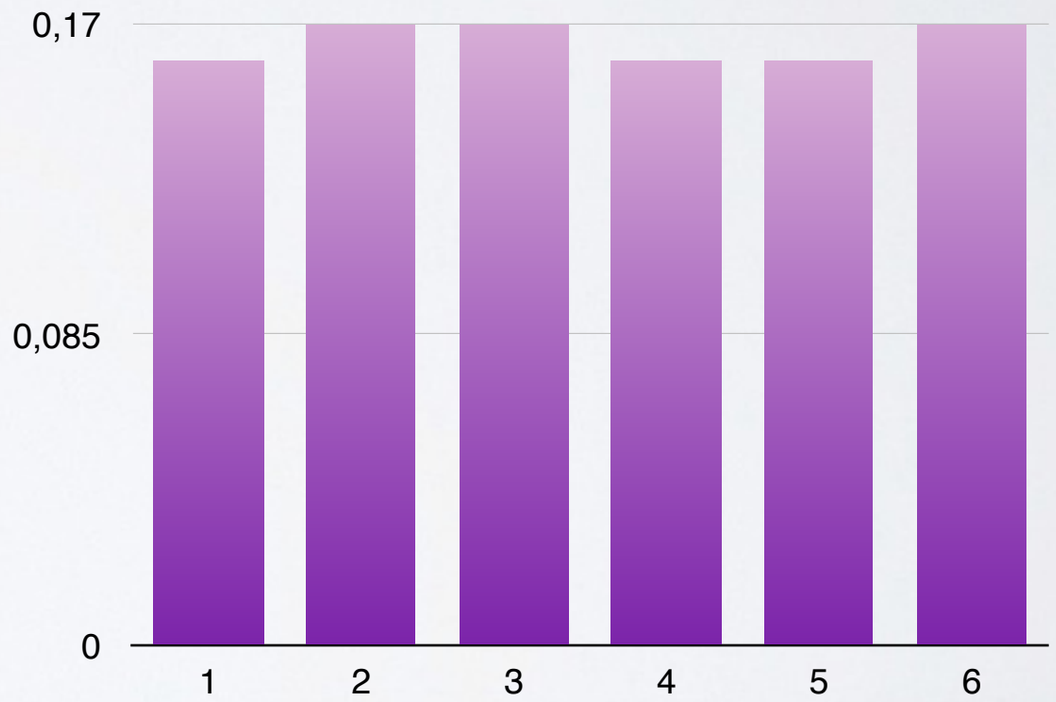
ROZKŁAD ZMIENNEJ LOSOWEJ DYSKRETNEJ

Rzut kostką do gry

Rzut kostką

Liczba oczek, X	$Pr(X)$
1	0.16
2	0.17
3	0.17
4	0.16
5	0.16
6	0.17
Suma	1.00

Rozkład równomierny



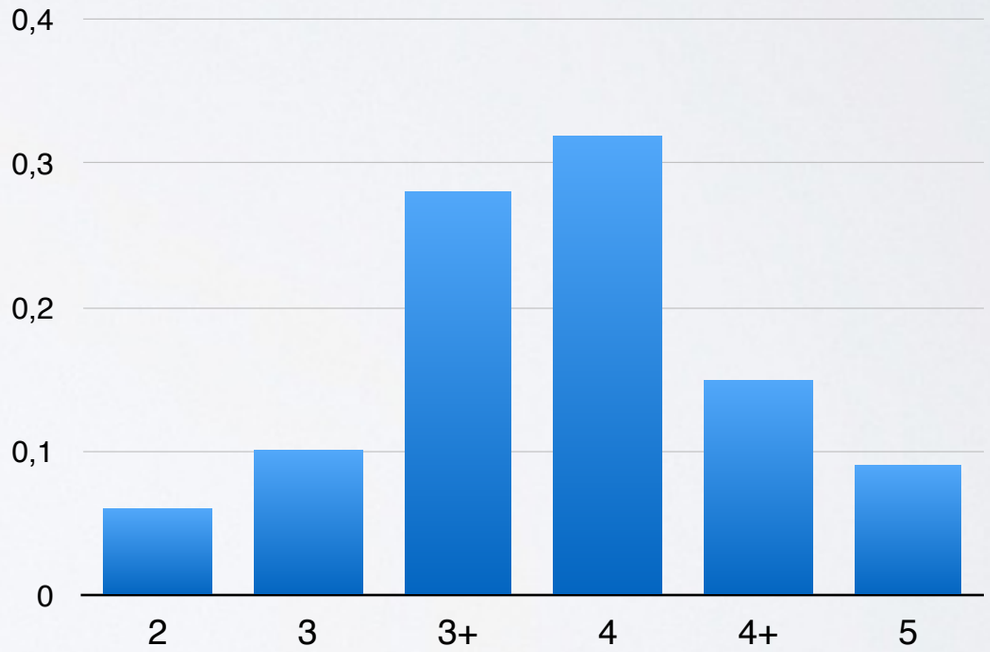
ROZKŁAD ZMIENNEJ LOSOWEJ DYSKRETNEJ

Oceny z kolokwium

Wartości skupione między 3+ a 4

Oceny z kolokwium

Ocena, X	$Pr(X)$
2	0.06
3	0.10
3+	0.28
4	0.32
4+	0.15
5	0.09
Suma	1.00



PARAMETRY ROZKŁADU ZMIENNEJ DYSKRETNEJ

- Wartość oczekiwana, ***EX*** – wartość, wokół której skupiają się wartości zmiennej przy wielokrotnym powtarzaniu eksperymentu.
- Wariancja, σ^2 – miara rozproszenia wartości zmiennej wokół wartości oczekiwanej.
- Odchylenie standardowe, σ – charakteryzuje rozrzut zbioru wartości zmiennej wokół wartości oczekiwanej.

$$EX = \sum x_i P\{X = x_i\}$$

$$\sigma^2 = \sum (x_i - EX)^2 P\{X = x_i\} \qquad \sigma = \sqrt{\sigma^2}$$

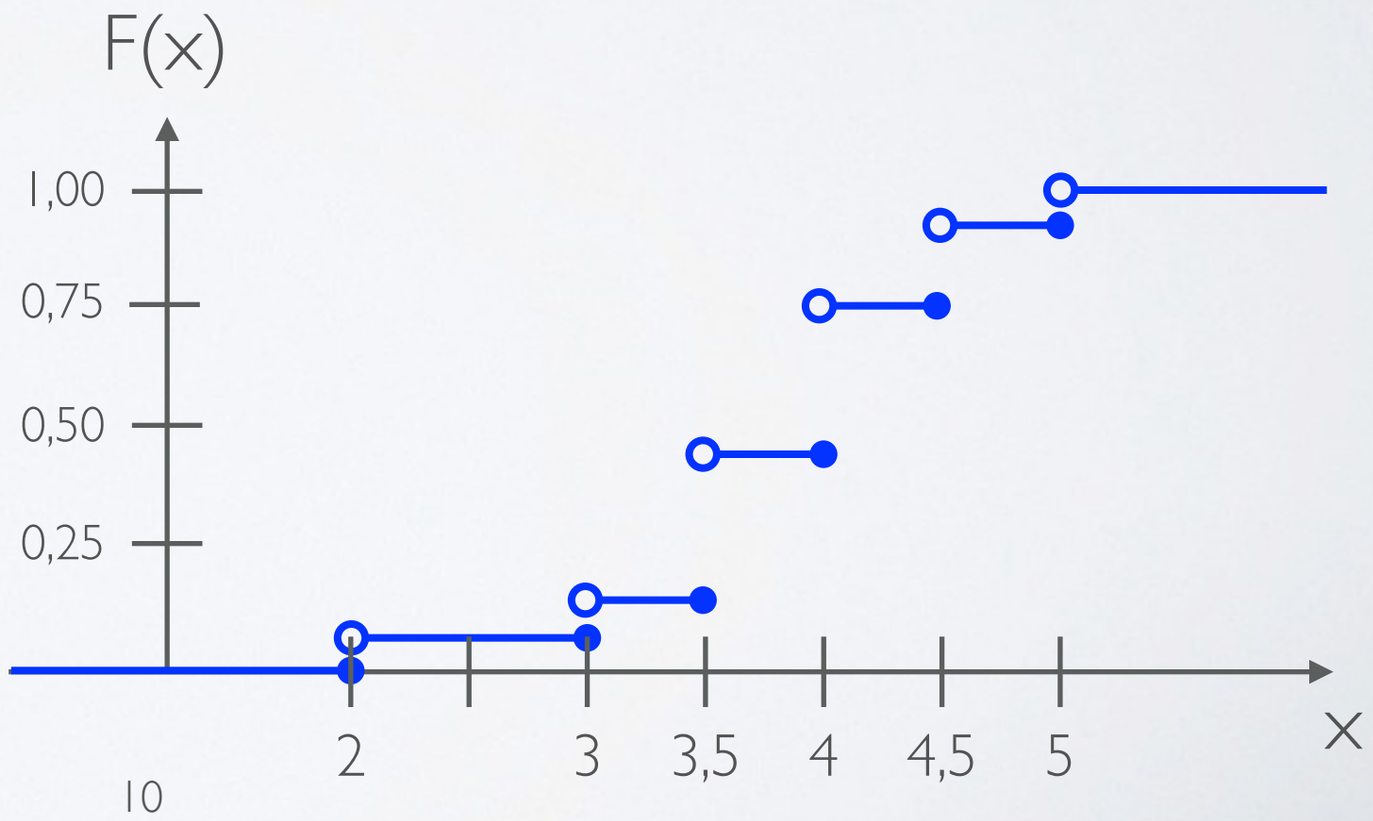
DYSTRYBUANTA ZMIENNEJ LOSOWEJ

- Dystrybuantą zmiennej losowej X nazywamy funkcję określoną na zbiorze liczb rzeczywistych, zdefiniowaną jako:

$$F(x) = P\{X < x\}, \quad x \in \mathcal{R}$$

Oceny z kolokwium

Ocena, X	Pr(X)	F(X)
2	0.06	0,06
3	0.10	0,16
3+	0.28	0,44
4	0.32	0,76
4+	0.15	0,91
5	0.09	1
Suma	1.00	



PARAMETRY ROZKŁADU ZMIENNEJ DYSKRETNEJ - CD.

- Kwantyl rzędu p zmiennej X jest to taka liczba x_p, że

$$F(x_p) = p$$

- Kwantylem rzędu p jest taka wartość x_p zmiennej losowej X, że wartości równe lub mniejsze od x_p przyjmowane są z prawdopodobieństwem co najmniej p.

PARAMETRY ROZKŁADU ZMIENNEJ DYSKRETNEJ - CD.

- Mediana – kwantyl rzędu 1/2
- Kwartyl – kwantyle rzędu 1/4, 2/4, 3/4
- Kwintyl – kwantyle rzędu 1/5, 2/5, 3/5, 4/5
- Percentyl – kwantyle rzędu 1/100, … 99/100

OBLICZANIE KWANTYLI – PRZYKŁAD

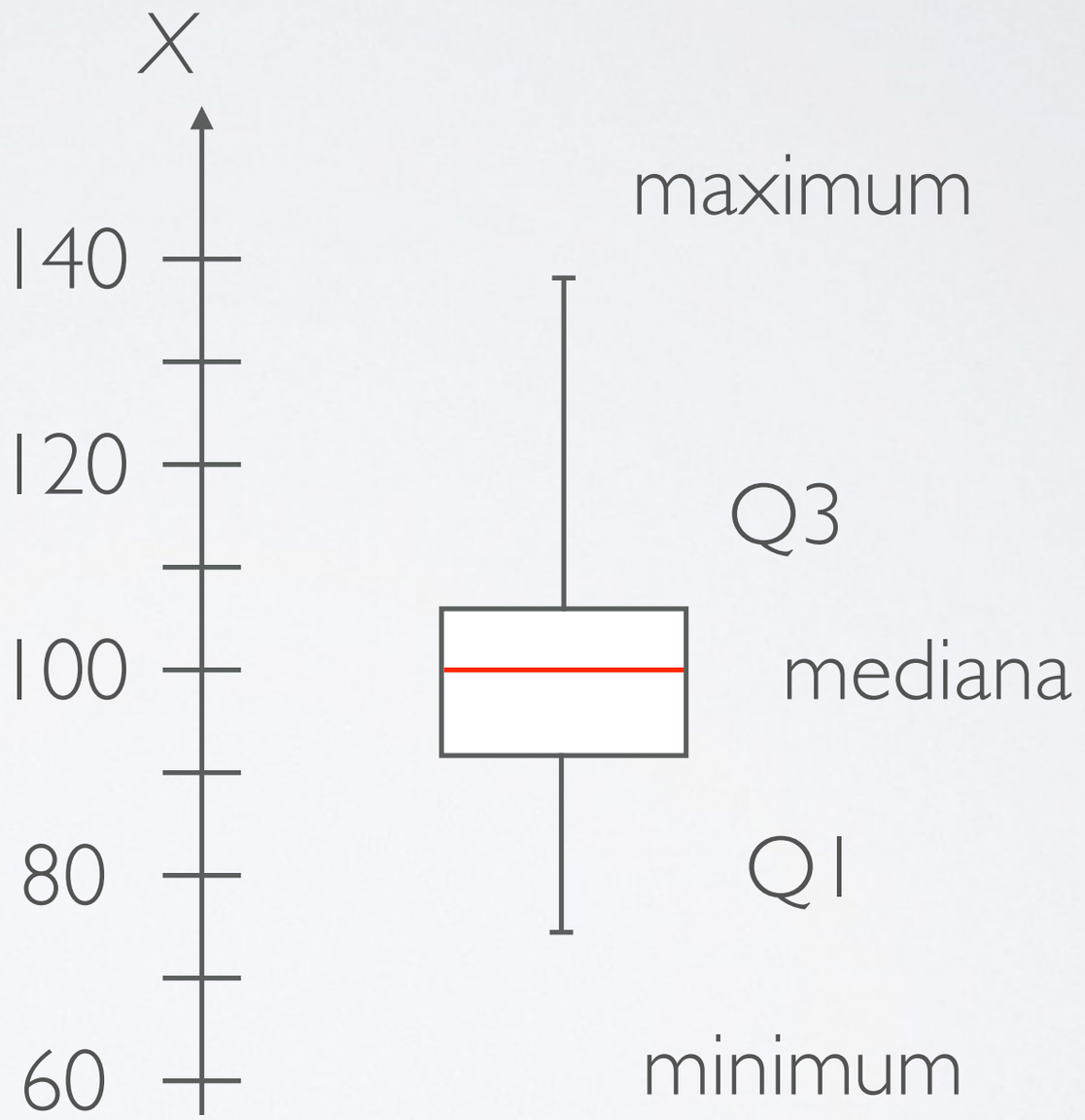
- Przypuśćmy, że zmierzono iloraz inteligencji u 20 osób.
- Wyniki uszeregowano w kolejności rosnącej:
  74, 80, 80, 85, **92**, 94, 97, 98, 98, **100**, 101, 101, 104, 104, **106**, 109, 112, 115, 128, 137.
- Kwantylem rzędu 0.25 jest liczba 92, gdyż 1/4 populacji ma wartość mniejszą lub równą 92.
- Kwantylem rzędu 0.75 jest liczba 106.

https://pl.wikipedia.org/wiki/Kwantyl

WYKRES PUDEŁKOWY

Iloraz inteligencji w próbie:

74, 80, 80, 85, **92**, 94, 97, 98, 98, **100**, 101, 101, 104, 104, **106**, 109, 112, 115, 128, 137



TYPY ROZKŁADÓW ZMIENNEJ LOSOWEJ DYSKRETNEJ

- Równomierny
- Zero-jedynkowy
- **Dwumianowy / Bernoulliego**
- Poissona
- Geometryczny

ROZKŁAD ZERO-JEDYNKOWY

- Rozkład zero-jedynkowy charakteryzuje się tym, że realizacja zmiennej losowej może przyjmować dwie wartości, tzn. 0 lub 1.

- Przy czym, Pr(X=1) = p, zaś Pr(X=0) = q = 1–p.

- Funkcja masy prawdopodobieństwa:

$$f(k;p) = p^k(1-p)^{1-k} \qquad k \in \{0,1\}$$

- $EX = p$, $\sigma^2 = p(1–p)$

ROZKŁAD DWUMIANOWY

- **Schemat Bernoulliego**: wykonujemy *n* razy doświadczenie, które może zakończyć się jednym z dwóch wyników: sukcesem lub porażką, z prawodopodbieństwami odpowiednio *p* oraz $q = 1 - p$.

- Zmienna losowa zdefiniowana jako **liczba sukcesów** w *n* próbach ma rozkład dwumianowy (Bernoulliego) określony wzorem:

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

$$k = 0, 1, \ldots, n$$

Jacob Bernoulli
1655-1705

ROZKŁAD DWUMIANOWY – PRZYKŁADY

- Według danych statystycznych (Diagnoza Społeczna 2013) liczba palaczy w Polsce wynosi ok. 26 proc. (8,5 mln osób)
- Załóżmy, że próba losowa wynosi 15 osób. Jakie jest prawdopodobieństwo, że 5 osób z tej próby pali papierosy?

$$\frac{15!}{5!(15-5)!} \cdot 0.26^5 \cdot (1-0.26)^{(15-5)} = 17.6\%$$

ROZKŁAD DWUMIANOWY – PRZYKŁAD #1

- Jakie jest prawdopodobieństwo, że w próbie będą dwie lub mniej osób palących?

$$P(x <= 2) = P(x = 2) + P(x = 1) + P(x = 0)$$

$$= 0.1416 + 0.0576 + 0.0109$$

$$= 21\%$$

```
>>> import scipy, scipy.stats
>>> pmf = scipy.stats.binom.pmf(x,n,p)
```

ROZKŁAD DWUMIANOWY – ZAŁOŻENIA

- Wzór Bernoulliego ma zastosowanie dla dowolnej zmiennej losowej, zdefiniowanej jako liczba sukcesów w próbie *n*-krotnej, przy czym
  - liczba n jest ustalona z góry,
  - prawdopodobieństwo sukcesu *p* nie zmienia się z próby na próbę,
  - próby są niezależne od siebie.
- Przy spełnieniu powyższych wymagań mówimy, że zmienna losowa ma rozkład dwumianowy.

KLUCZOWE PYTANIE #3

Dlaczego mówimy o rozkładzie prawdopodobieństwa zmiennej losowej?

ROZKŁAD DWUMIANOWY – PRZYKŁAD #2

- Rozkład dwumianowy ma kluczowe zastosowanie w analizie jednorodnej próby danych nominalnych

- Przykład: oszacowanie szans przeżycia dzieci przedwcześnie urodzonych (John Hopkins Hospital, Baltimore, Maryland, USA)

- Próba: 39 noworodków urodzonych w 25 tygodniu ciąży. 31 dzieci dożyło co najmniej 6 miesiąca.

- Wszystkie dzieci były poddawane tym samym zabiegom podtrzymującym życie. W przeciwnym przypadku byłaby to próba mieszana.

ROZKŁAD DWUMIANOWY – PRZYKŁAD #2

- Realizacja zmiennej losowej przyjmuje dwie wartości – dziecko dożyje co najmniej do 6-ego miesiąca albo nie.

- Z obserwacji danych historycznych wynika, że prawdopodobieństwo przeżycia wynosi 31/39 = 79,5%.

- Stawiamy pytanie, jaka jest rzeczywista szansa na przeżycie całej populacji? Chcemy mieć określoną pewność, że ten wskaźnik nie będzie znacząco odbiegał od wartości szacunkowej. W tym celu posługujemy się pojęciem przedziału ufności.

- Załóżmy, że chcemy z 95% pewnością znać przedział, w którym mieści się prawdziwa wartość prawdopodobieństwa przeżycia.

ROZKŁAD DWUMIANOWY – PRZYKŁAD #2

- Niezależnie od próby, liczba noworodków, która dożyje co najmniej 6-ego miesiąca po porodzie, będzie miała rozkład dwumianowy.

- Przedział ufności oznaczamy jako (p_L, p_U).

- Poziom ufności 95% oznacza, że prawdopodobieństwo p_L uzyskania wartości estymowanej (79,5%) lub większej nie może być mniejsze niż 2,5%, a prawdopodobieństwo p_U uzyskania wartości estymowanej lub mniejszej nie będzie większe niż 2,5%.

- Na początku spróbujmy wyznaczyć p_L posługując się wyłącznie intuicją…

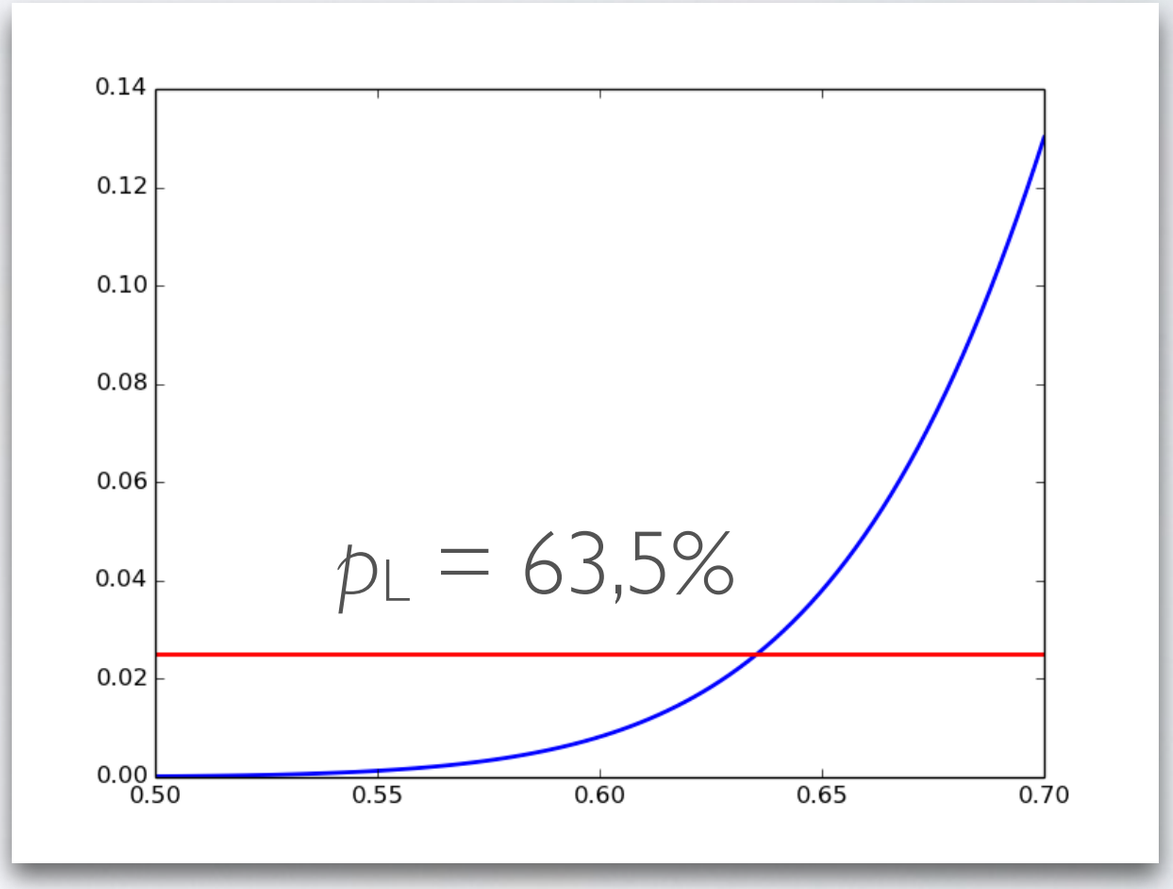
ROZKŁAD DWUMIANOWY – PRZYKŁAD #2

- Niech $p_L = 50\%$.
- Wówczas prawdopodobieństwo tego, że z próby 39 noworodków-wcześniaków przeżyje co najmniej 31 dzieci wynosi 0,015%.

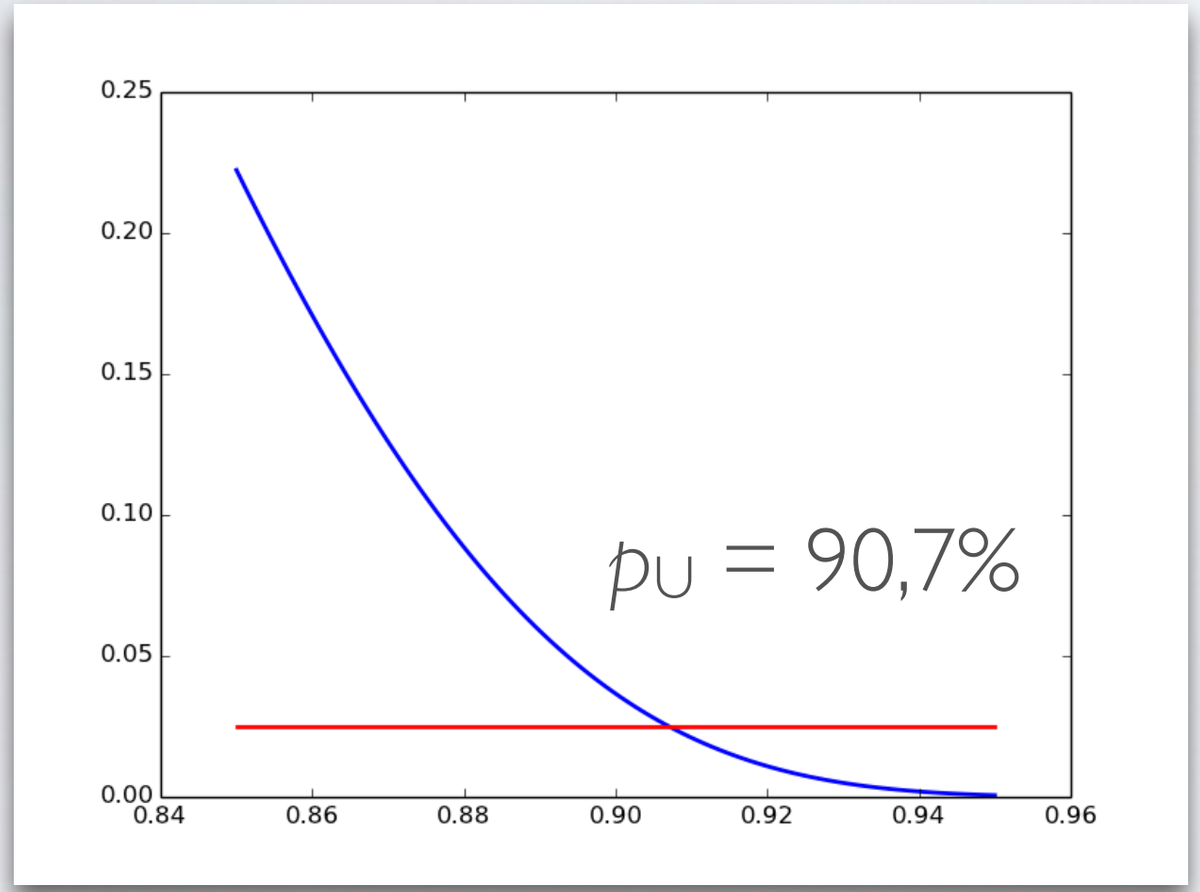
$$
\begin{aligned}
P(x \geq 31) &= P(x = 31) + P(x = 32) + \dots \\
&= \tfrac{39!}{31!8!} \cdot 0.5^{31}(1 - 0.5)^{8} + \dots \\
&= 0.00015
\end{aligned}
$$

ROZKŁAD DWUMIANOWY – PRZYKŁAD #2

```
import scipy, scipy.stats, numpy
n = 39
k = scipy.linspace(31, 39, 9)
pL = scipy.linspace(0.5, 0.7, 100)

bound = numpy.zeros(100) + 0.025
pmf = numpy.zeros(100)

for i in range(0, 100):
    pmf[i] = sum(scipy.stats.binom.pmf(k, n,
pL[i]))

import matplotlib.pyplot as plt
plt.plot(pL, pmf, 'b', pL,
bound,'r',linewidth=2)
plt.show()
```



ROZKŁAD DWUMIANOWY – PRZYKŁAD #2

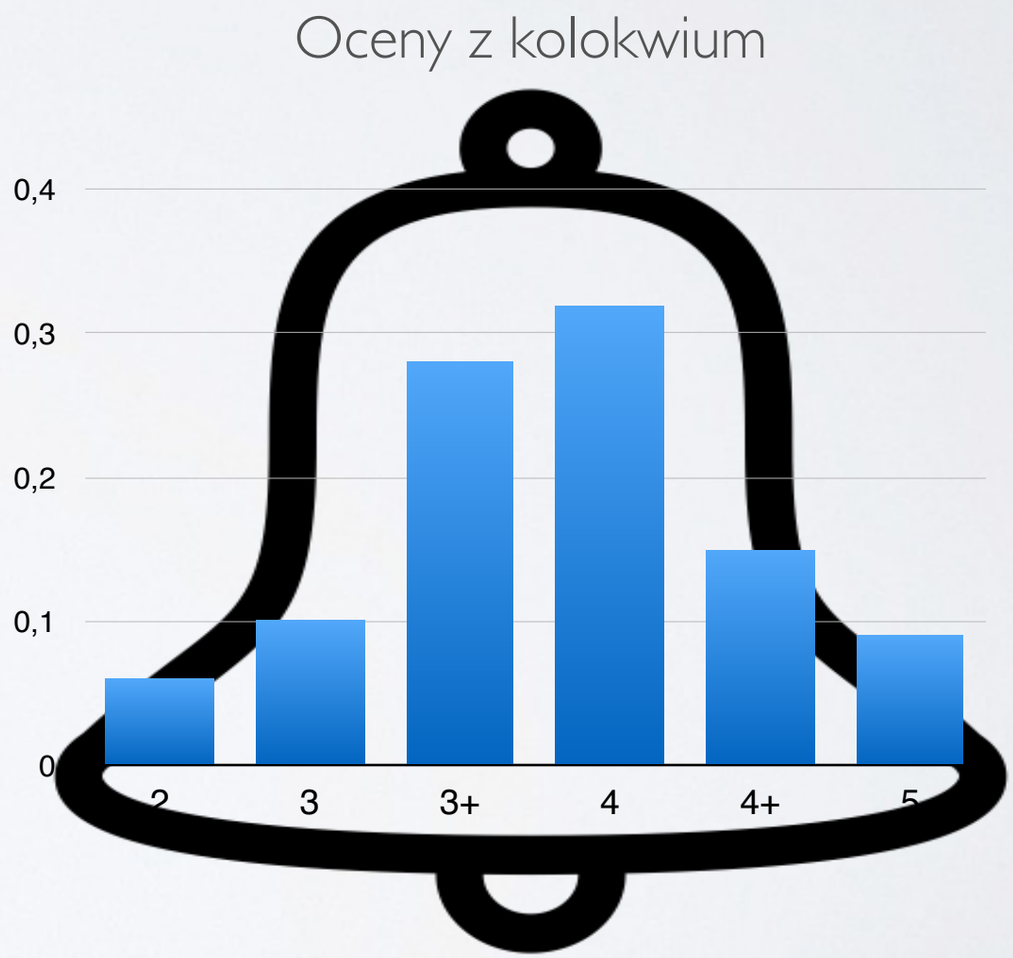



```
…

n = 39
k = scipy.linspace(0, 31, 32)

…
```

Na poziomie ufności 95% prawdopodobieństwo przeżycia noworodka (dla całej poplacji) urodzonego w 25 tygodniu ciąży mieści się w przedziale (63,5%; 90,7%).

ROZKŁADY ZMIENNEJ LOSOWEJ CIĄGŁEJ

- Jednostajny
- Wykładniczy
- **Normalny / Gaussa**
- t-Studenta
- χ-kwadrat



ROZKŁAD NORMALNY – ODKRYWCY

Abraham de Moivre
1667-1754

Carl Friedrich Gauß
1777-1855

Pierre-Simon Laplace
1749-1827

ROZKŁAD NORMALNY

- Obserwacja wielu zjawisk w przyrodzie pozwala stwierdzić, że odbywają się one zgodnie z rozkładem normalnym, lub bardzo zbliżonym do niego.

- Wynika to z **centralnego twierdzenia granicznego**, które stanowi, że suma dużej liczby zmiennych losowych o dowolnym, takim samym rozkładzie zbliża się do rozkładu normalnego.

- Parametrami rozkładu normalnego jest wartość średnia i odchylenie standardowe.

- Funkcja gęstości prawdopodobieństwa wyraża się wzorem:

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

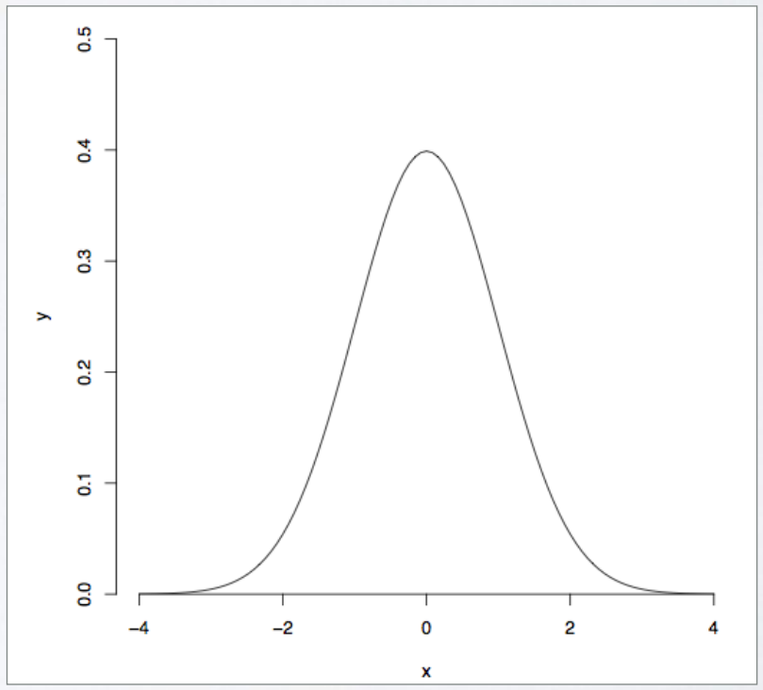
ROZKŁAD NORMALNY

- Jeżeli zmienna X ma rozkład $N(\mu,\sigma)$, to zmienna $Z=(X-\mu)/\sigma$ ma rozkład $N(0,1)$ zwany standardowym rozkładem normalnym.

- Gęstość i dsytrybuanta tego rozkładu wyrażają się wzorami:

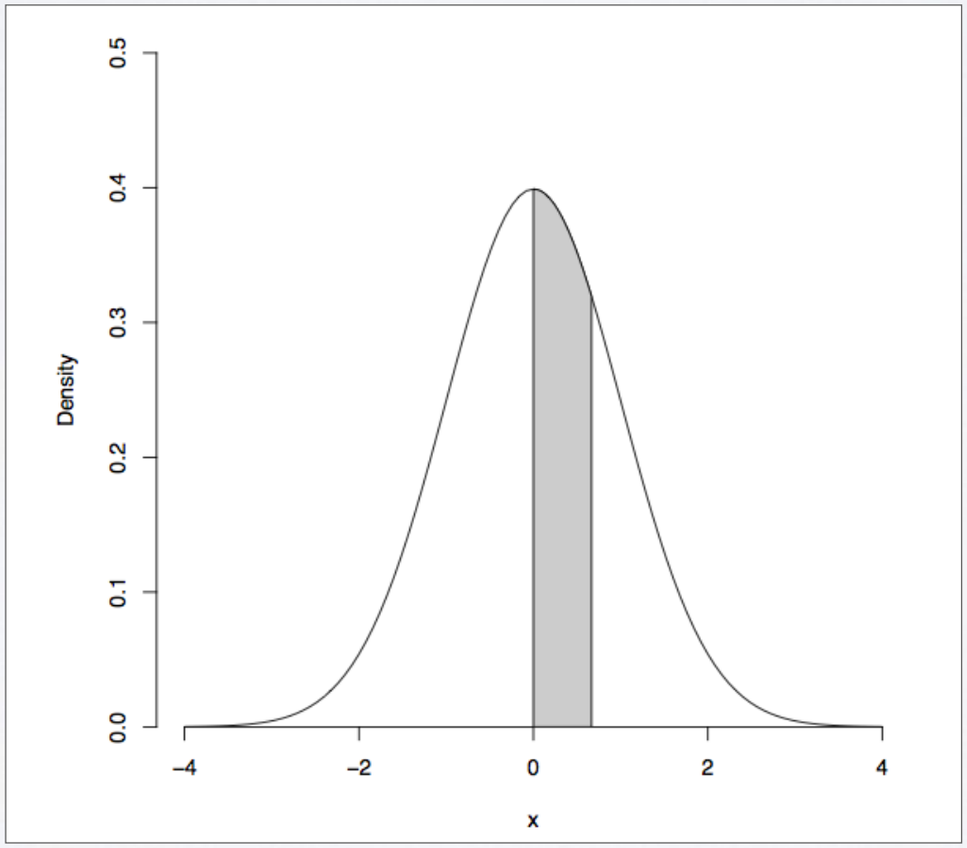
$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$$

$$F(x) = \int_{-\infty}^{x} f(t)dt$$



ROZKŁAD NORMALNY

- Prawdopodobieństwo można wyznaczyć jako powierzchnię pod krzywą normalną:



ROZKŁAD NORMALNY

- Dystrybuanta $F(x)$ standardowego rozkładu normalnego dla $x \geq 0$ jest stablicowana. Dla $x < 0$ zachodzi $F(x) = 1 - F(-x)$.

- Aby wyznaczyć prawdopodobieństwo, że zmienna X, która ma rozkład normalny przyjmie wartość z przedziału (a,b) można skorzystać ze wzoru:

$$P\{X \in (a,b)\} = P\left\{Z \in \left(\frac{a-\mu}{\sigma}, \frac{b-\mu}{\sigma}\right)\right\} = F\left(\frac{a-\mu}{\sigma}\right) - F\left(\frac{b-\mu}{\sigma}\right)$$

ROZKŁAD NORMALNY

Tablica 2. Dystrubuanta $F(x)$ rozkładu normalnego $N(0,1)$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	50000	50399	50798	51197	51595	51994	52392	52790	53188	53586
0.1	53983	54380	54776	55172	55567	55962	56356	56749	57142	57535
0.2	57926	58317	58706	59095	59483	59871	60257	60642	61026	61409
0.3	61791	62172	62552	62930	63307	63683	64058	64431	64803	65173
0.4	65542	65910	66276	66640	67003	67364	67724	68082	68439	68793
0.5	69146	69497	69847	70194	70540	70884	71226	71566	71904	72240
0.6	72575	72907	73237	73565	73891	74215	74537	74857	75175	75490
0.7	75804	76115	76424	76730	77035	77337	77637	77935	78230	78524
0.8	78814	79103	79389	79673	79955	80234	80511	80785	81057	81327
0.9	81594	81859	82121	82381	82639	82894	83147	83398	83646	83891
1.0	84134	84375	84614	84849	85083	85314	85543	85769	85993	86214
1.1	86433	86650	86864	87076	87286	87493	87698	87900	88100	88298
1.2	88493	88686	88877	89065	89251	89435	89617	89796	89973	90147
1.3	90320	90490	90658	90824	90988	91149	91308	91466	91621	91774
1.4	91924	92073	92220	92364	92507	92647	92785	92922	93056	93189

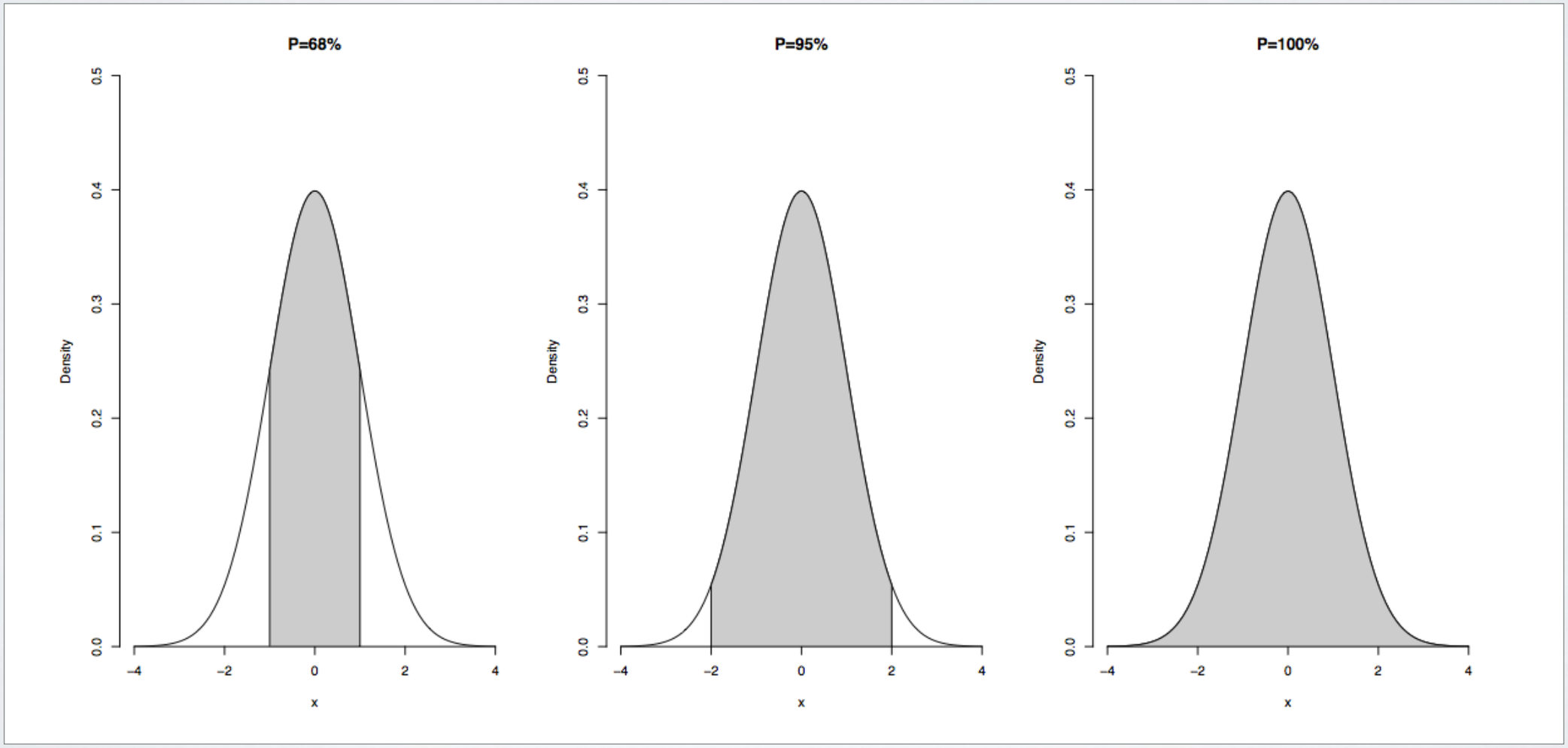
ROZKŁAD NORMALNY – WYZNACZANIE PERCENTYLI

- Jaka jest wartość 60-tego percentyla w rozkładzie normalnym?
- Ups... W zasadzie nie ma takiej wartości w tablicy.
- Gdzieś pomiędzy 0,25 a 0,26 (a tak naprawdę 0,2533) .

z	**0,00**	0,01	0,02	**0,03**	0,04	0,05	**0,06**	0,07
0,0	**0,5000**	0,5040	0,5080	**0,5120**	0,5160	0,5199	**0,5239**	0,5279
0,1	**0,5398**	0,5438	0,5478	**0,5517**	0,5557	0,5596	**0,5636**	0,5675
0,2	**0,5793**	0,5832	0,5871	**0,5910**	0,5948	0,5987	**0,6026**	0,6064
0,3	**0,6179**	0,6217	0,6255	**0,6293**	0,6331	0,6368	**0,6406**	0,6443
0,4	**0,6554**	0,6591	0,6628	**0,6664**	0,6700	0,6736	**0,6772**	0,6808
0,5	**0,6915**	0,6950	0,6985	**0,7019**	0,7054	0,7088	**0,7123**	0,7157
0,6	**0,7257**	0,7291	0,7324	**0,7357**	0,7389	0,7422	**0,7454**	0,7486
0,7	**0,7580**	0,7611	0,7642	**0,7673**	0,7704	0,7734	**0,7764**	0,7794

ROZKŁAD NORMALNY

- Dla zmiennej losowej o rozkładzie $N(\mu,\sigma)$ obowiązuje **prawo trzech sigm**:



ROZKŁAD NORMALNY – PRZYKŁAD

- Doświadczenie polega na badaniu czasu snu pacjentów chorych na pewną chorobę.
- Przyjmujemy, że czas snu ma rozkład normalny ze średnią 500 minut i odchyleniem standardowym 20 minut.
- Długość snu uważa się za prawidłową, jeżeli mieści się w przedziale (460, 520).
- Jaki jest odsetek pacjentów o prawidłowej długości snu?

ROZKŁAD NORMALNY – PRZYKŁAD

- Wykonujemy obliczenia według wzoru:

$$
\begin{aligned}
P\{X \in (460, 520)\} &= F\left(\tfrac{520-500}{20}\right) - F\left(\tfrac{460-500}{20}\right) \\
&= F(1) - F(-2) \\
&= F(1) - (1 - F(2)) = F(2) + F(1) - 1 \\
&= 0.97725 + 0.84134 - 1 = 0.81859.
\end{aligned}
$$

Odsetek pacjentów o prawidłowej długości snu wynosi 82%.

ROZKŁAD NORMALNY – PYTHON

```
>>> import matplotlib.pyplot as plt
>>> import numpy as np
>>> mu, sigma = 0, 0.1
>>> s = np.random.normal(mu, sigma, 1000)
>>> count, bins, ignored = plt.hist(s, 30, normed=True)
>>> plt.plot(bins, 1/(sigma * np.sqrt(2 * np.pi)) *
...                np.exp( - (bins - mu)**2 / (2 * sigma**2) ),
...          linewidth=2, color='r')
>>> plt.show()
```