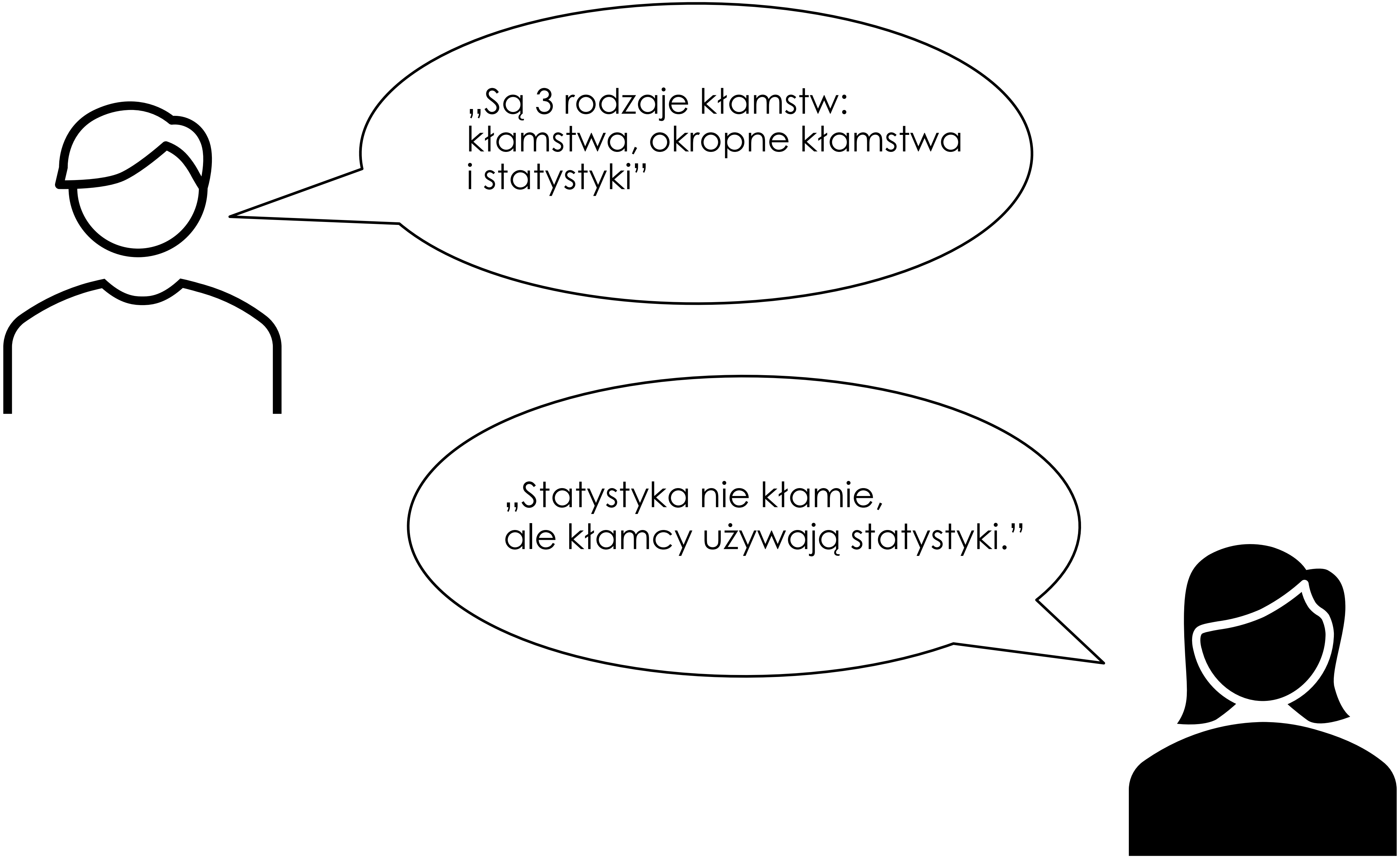
Model statystyczny, estymacja - podsumowanie

PAWEŁ PIASECKI


„Są 3 rodzaje kłamstw: kłamstwa, okropne kłamstwa i statystyki”
„Statystyka nie kłamie, ale kłamcy używają statystyki.”

Statystyka opisowa
vs
statystyka matematyczna

Statystyka opisowa (laboratoria 4) – liczyliśmy różne statystyki (średnią, odchylenie standardowe, medianę, kurtozę, skośność) i wyciągaliśmy wnioski dotyczące badanego zbioru (**bez ich generalizowania** na wszystkie możliwe dane z danego procesu/badania !!!)

Aby móc generalizować i uogólniać wnioski musimy w jakiś sposób modelować **dane, których nie widzimy.**

Naturalnym aparatem do takiego modelowania jest **rachunek prawdopodobieństwa**. Możemy dzięki niemu powiedzieć, że nasze dane pochodzą z jakiegoś **rozkładu prawdopodobieństwa**. Przy pomocy **estymacji parametrów takiego rozkładu** możemy zdobyć informację o całej populacji.

Statystyka matematyczna – zajmuje się analizą zjawisk masowych przy użyciu metod rachunku prawopodobieństwa.

Problem

- Chcemy produkować biurka dla informatyków.
- Jakiej wysokości powinny być te biurka? Jak to policzyć?

Elementy populacji

Język informatyki →	atrybuty	obiekt
Język statystyki →	cechy/zmienne	element populacji

IQ

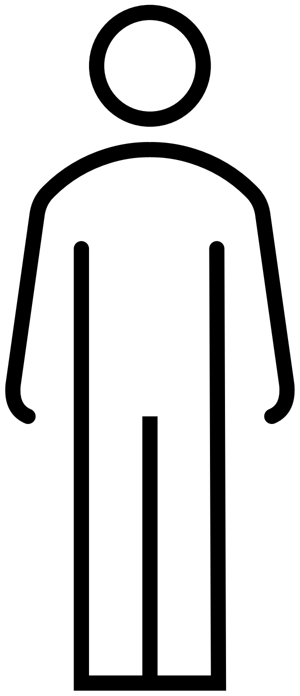
wada wzroku

puls spoczynkowy

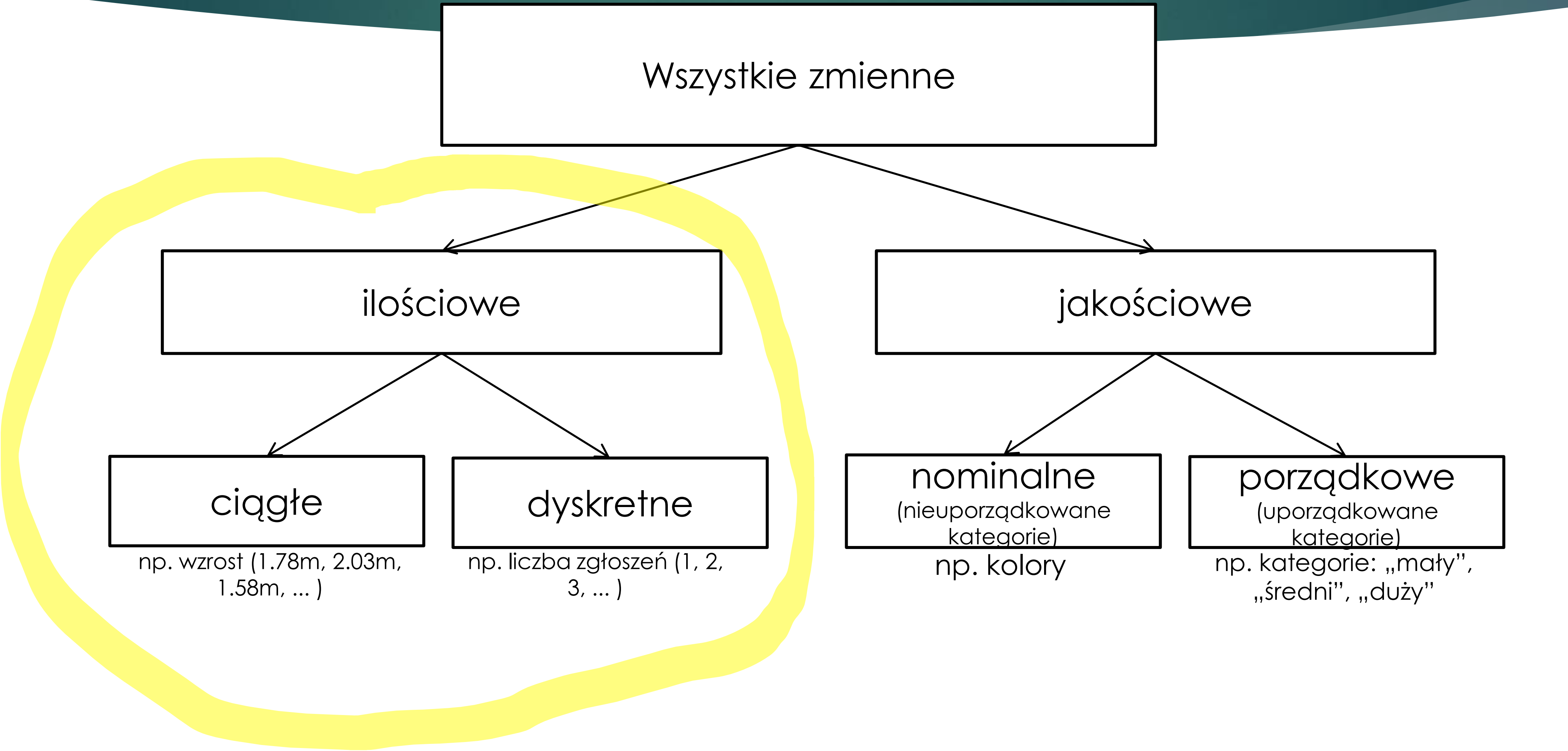
obwód bicepsa

wzrost

rozmiar buta



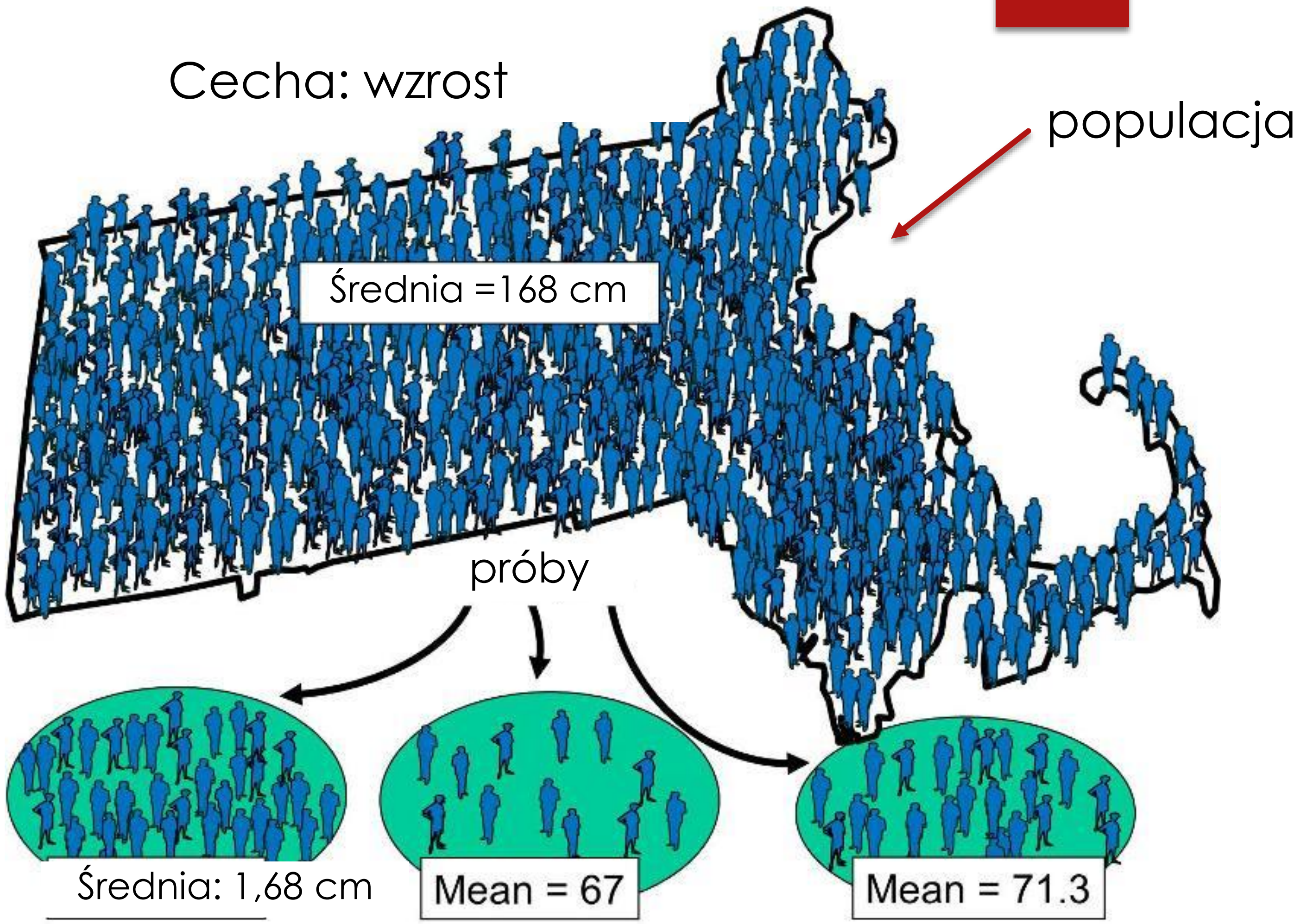
Typy cech statystycznych



Populacja vs próba

- Populacja – zbiór elementów podlegających badaniu lub analizie.

- Próba – podzbiór populacji. Elementy próby nazywamy obserwacjami.



Źródło: http://sphweb.bumc.bu.edu/otlt/MPHModules/BS/BS704_BiostatisticsBasics/BS704_BiostatisticsBasics2.html

Problem - przypomnienie

- Chcemy produkować biurka dla informatyków.
- Jakiej wysokości powinny być te biurka? Jak to policzyć?

- Zmierzyć wszystkich (całą populację)? **NIE**

- Wziąć próbę z populacji? **TAK**

Uwaga: Jeżeli mamy do dyspozycji całą populację, to nie potrzebujemy statystyki!!!

Przykład: Chcemy produkować/zamówić biurka dla pracowników naszej firmy.

Jak wybrać próbę?

Przykład zły: sondaż wyborczy, wybieramy osoby z zachodu Polski (najprawdopodobniej wygrałoby PO)

Przykład dobry: sondaż wyborczy, wybieramy osoby „rozsiane” po całej Polsce

próba prosta
=
próba reprezentatywna

Próba prosta – próbą prostą n-wymiarową z rozkłądu prawdopodobieństwa o dystrybuancie F nazywamy ciąg niezależnych zmiennych losowych $X_1, X_2, \ldots, X_n$ o jednakowym rozkładzie, $\forall_{i \in \{1,2,\ldots,n\}} P(X_i \leq x) = F(x)$.

Cel: znalezienie rozkładu **populacji** na podstawie **próby**

Zawsze zakładamy, że **populacja** posiada pewien **rozkład**, który chcemy znaleźć.

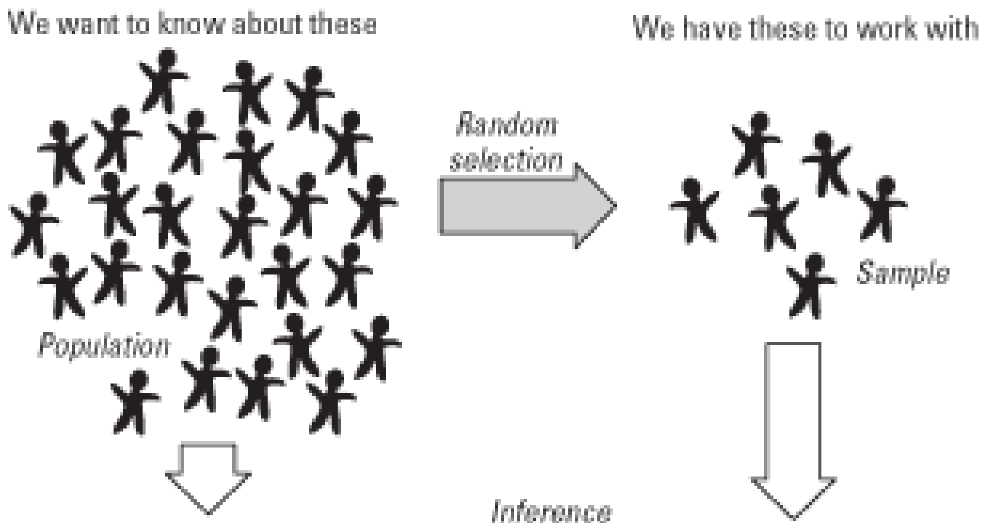
Znaleźć rozkład =

= określić **typ** (np. normalny, poissona) + wyestymować (policzyć wartość) **parametry** =

= określić **wzór** na prawdopodobieństwo określonych wartości, przykłady (kolejne 2 slajdy)

Przykład 1

Badana cecha: wzrost
Zakładany rozkład dla wzrostu – **rozkład normalny**



Parametry: μ, σ

Estymator parametru μ : $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Estymator parametry σ : $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$

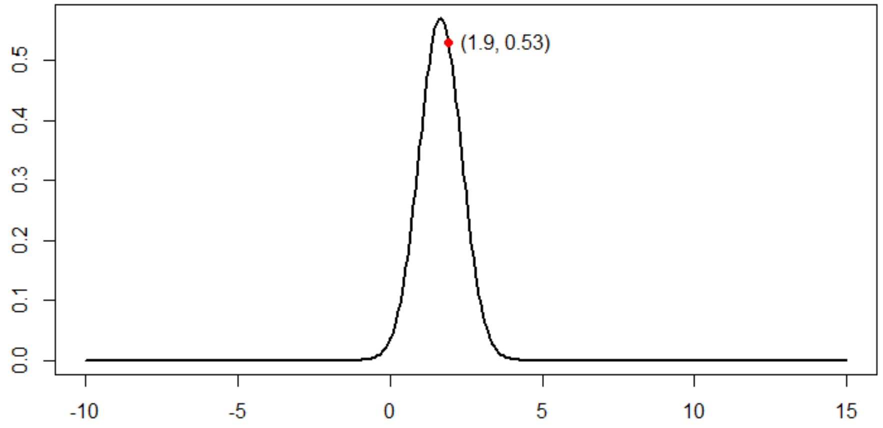
Gęstość rozkładu normalnego

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

Gęstość wystymowanego rozkładu normalnego
($\hat{\mu} = 1.65, \hat{\sigma} = 0.7$)

$$\boldsymbol{f_{1.65,0.7}(x) = \frac{1}{0.7\sqrt{2\pi}} exp\left(\frac{-(x-1.65)^2}{2 * 0.7^2}\right)},$$

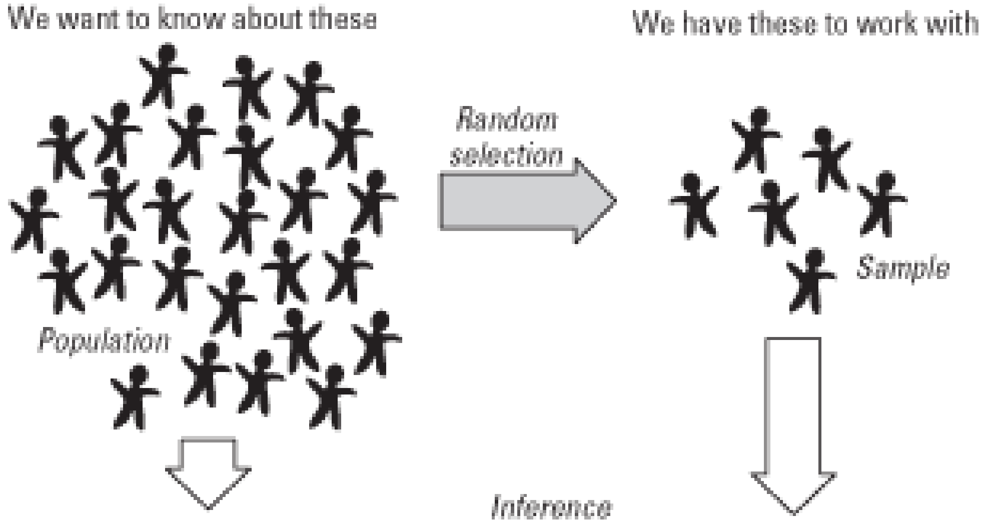
np. $f_{\mathbf{1.65,0.7}}(1{,}9) = 0.53$,
tzn. gęstość prawdopodobieństwa tego, że osoba z popuacji ma 1,9 m wzrostu jest równa 0.53



Źródło grafiki: https://www.cliffsnotes.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics

Przykład 2

Badana cecha: wzrost
Zakładany rozkład dla wzrostu – **rozkład wykładniczy**



Parametr: λ

Estymator parametru λ : $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$

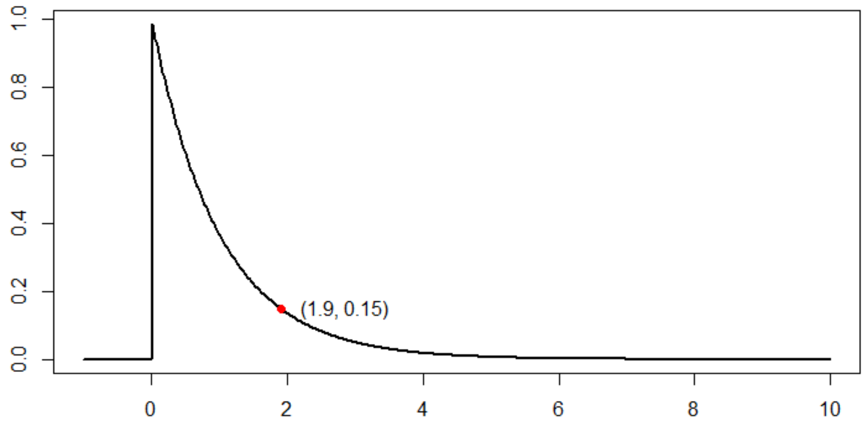
Gęstość rozkładu wykładniczego

$$f_\lambda(x) = \lambda e^{-\lambda x}$$

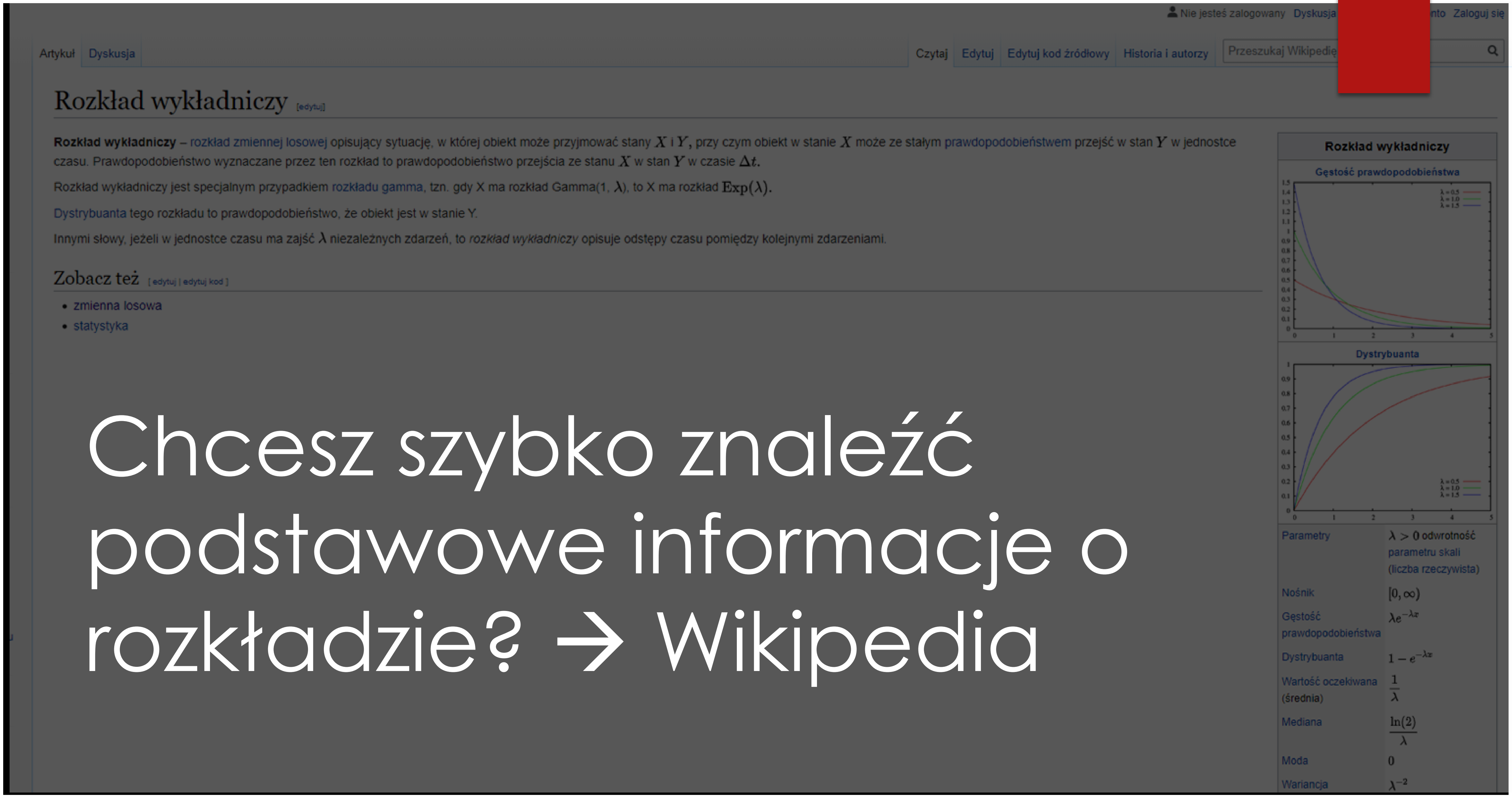
Gęstość wystymowanego rozkładu normalnego

$$f_1(x) = 1 \ * e^{-1*x},$$

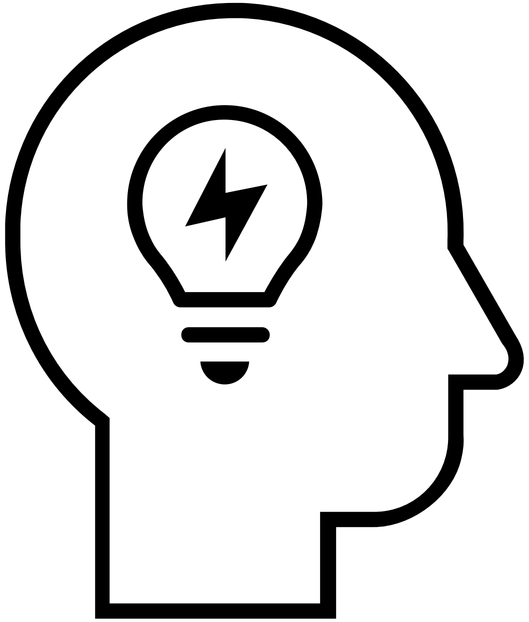
np. $f_1(1{,}8) = 0,16,$
tzn. gęstość prawdopodobieństwa tego, że osoba z popuacji ma 1,9 m wzrostu jest równa 0.15



Źródło grafiki: https://www.cliffsnotes.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics



Chcesz szybko znaleźć podstawowe informacje o rozkładzie? → Wikipedia



CZAS NA ROZWIĄZANIE PROBLEMU Z WYSOKOŚCIĄ BIUREK!

Eksploracja danych

Zbiór danych: 274 kobiety i 274 mężczyzn wybranych losowo wśród informatyków.

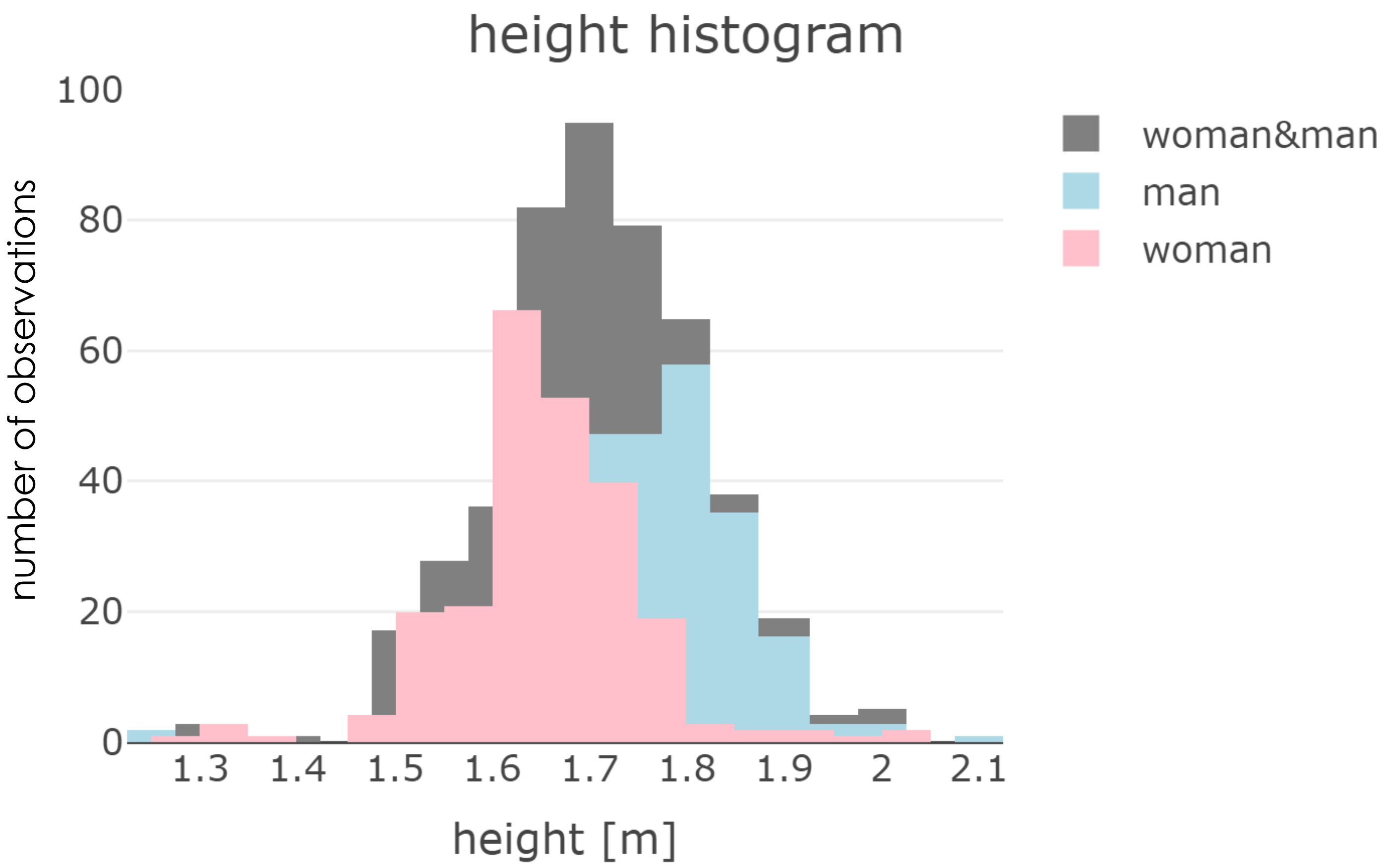
```
> heights %>% select(sex, height) %>% head()
     sex height
1   Male 1.9050
2   Male 1.7780
3   Male 1.7272
4   Male 1.8796
5   Male 1.5494
6 Female 1.6510
```

Załóżmy, że chcemy mieć 3 rodzaje biurek:

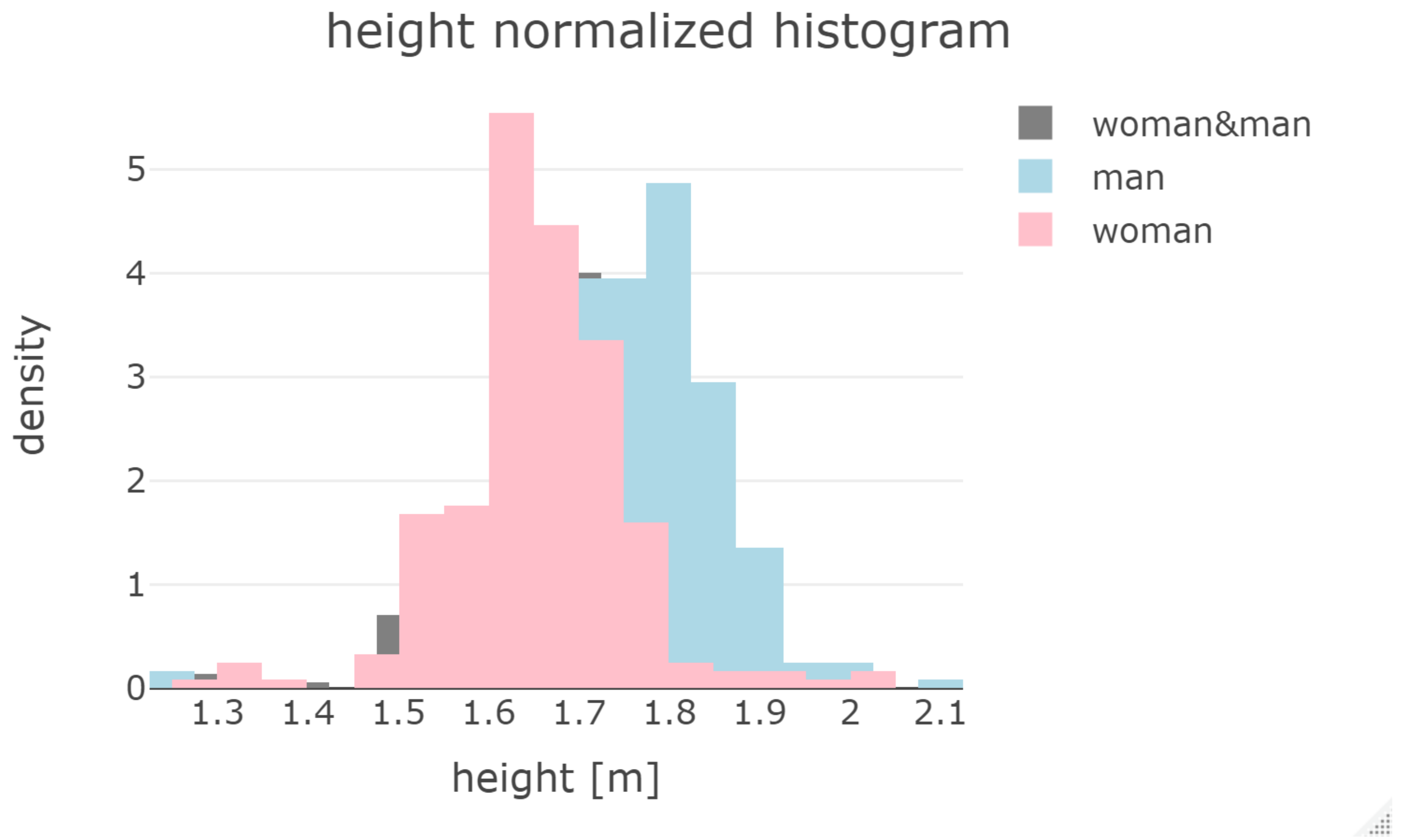
1. Różowe kobiece
2. Niebieskie męskie
3. Szare unisex

Pytanie z serii rocket science:

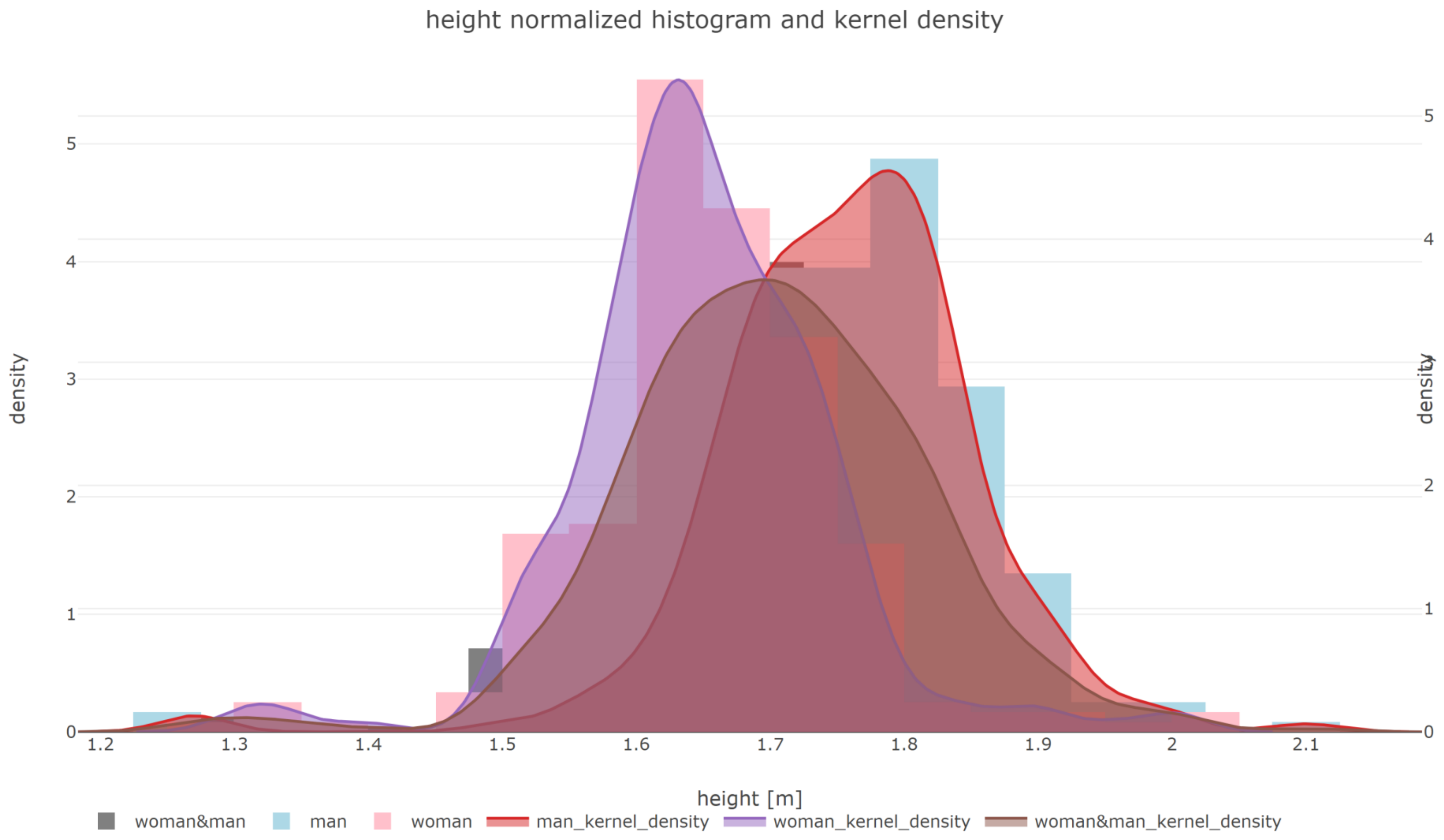
Jakie rozkład tutaj dopasować?



Znormalizowany histogram → suma pól słupków = 1

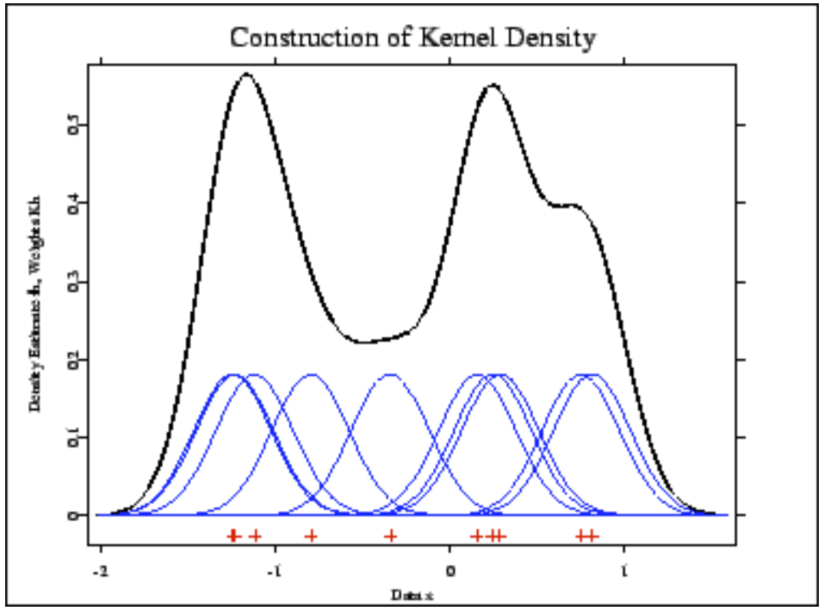


Dodajmy teraz jądrowy estymator gęstości



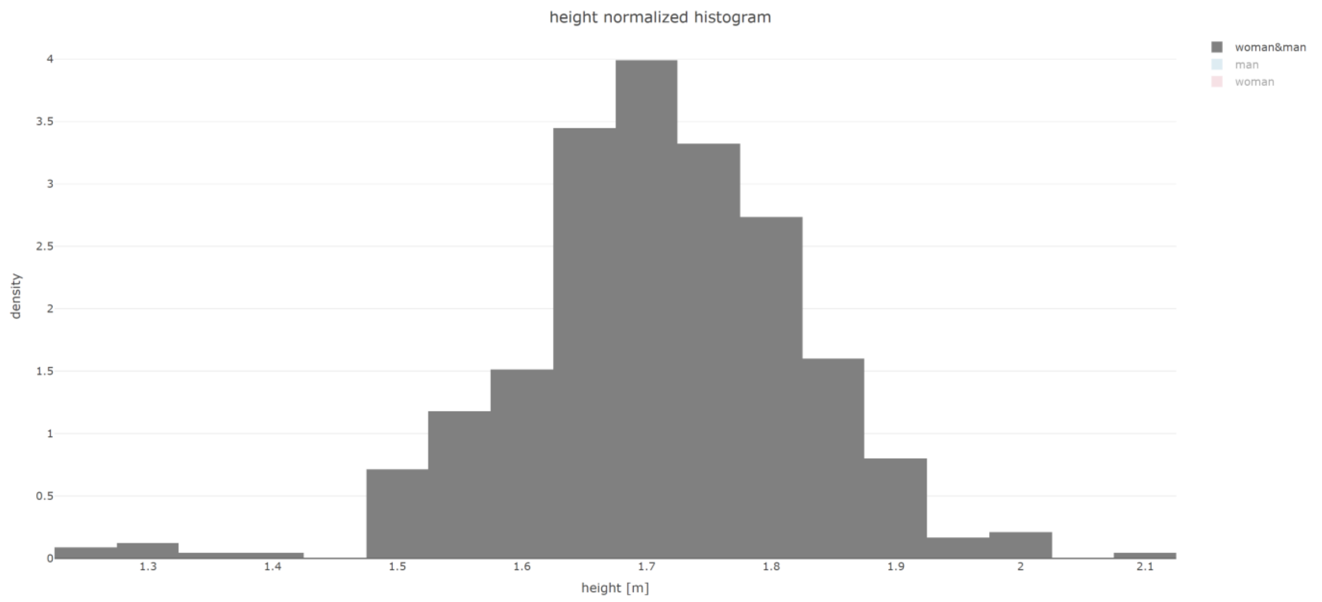
Jak tworzona jest gęstość jądrowa?
(dodatek dla chętnych)

Kernel density plots are a way of smoothing the distribution into a line, rather than bars, allowing for continuity. A kernel is a line shape which can be described using a mathematical function. Kernel functions “fill in the gaps”. It does this by applying a kernel to every data point. Look at the graph below. + indicates a data point, and directly above each point is the peak of a gaussian bell curve, which is an example of a kernel.



Źródło: http://heathershawresearch.com/2018/07/22/kernel-density-estimations-vs-histograms/

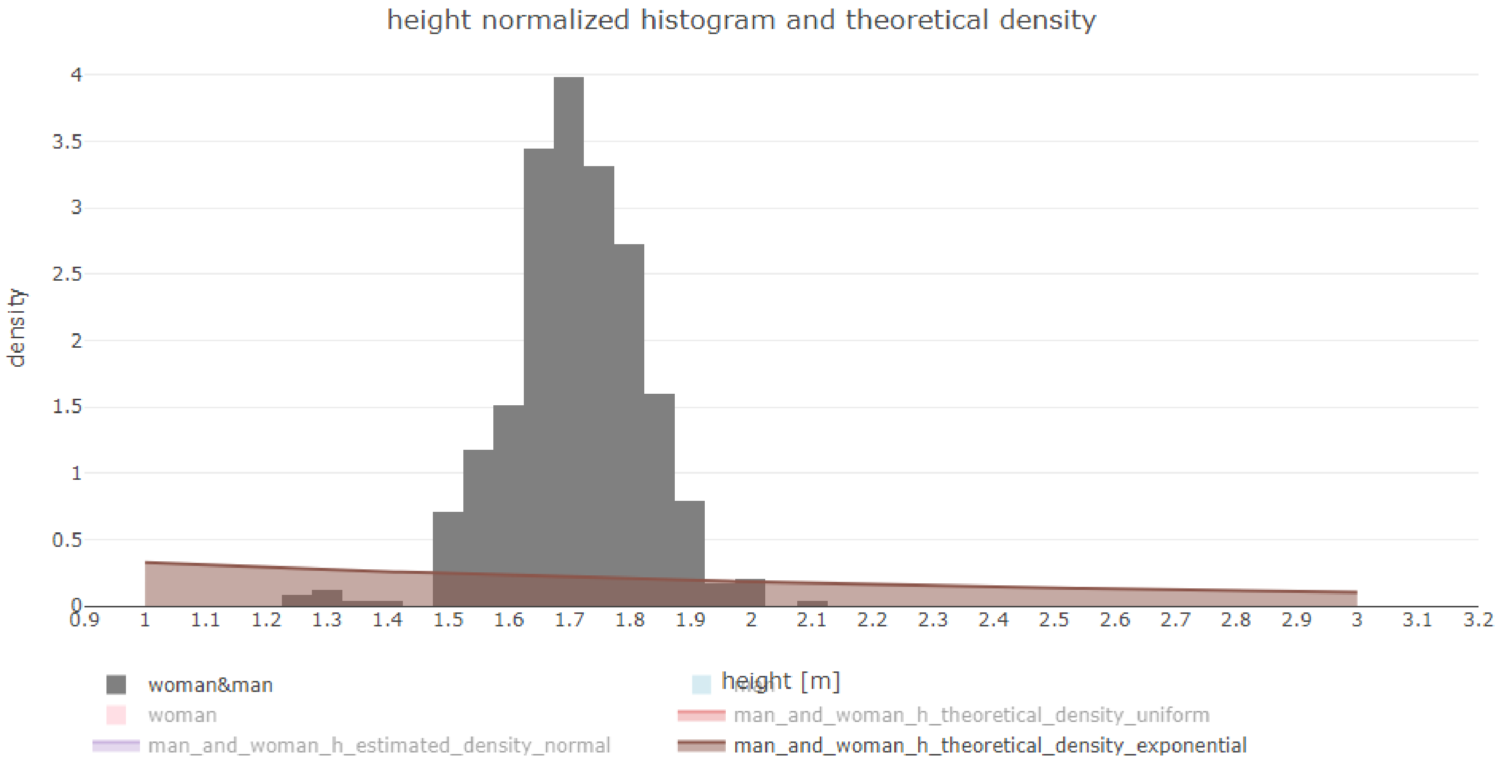
No dobrze, ale... Ciągle nie znamy **rozkładu teoretycznego**, tzn. rozkładu zakładanego dla całej populacji/dla wszystich kobiet/dla wszystkim meżczyzn.



Zajmijmy się teraz tylko rozkłaem wzrostu **wszystkich informatyków**. Jaki rozkład dopasować?

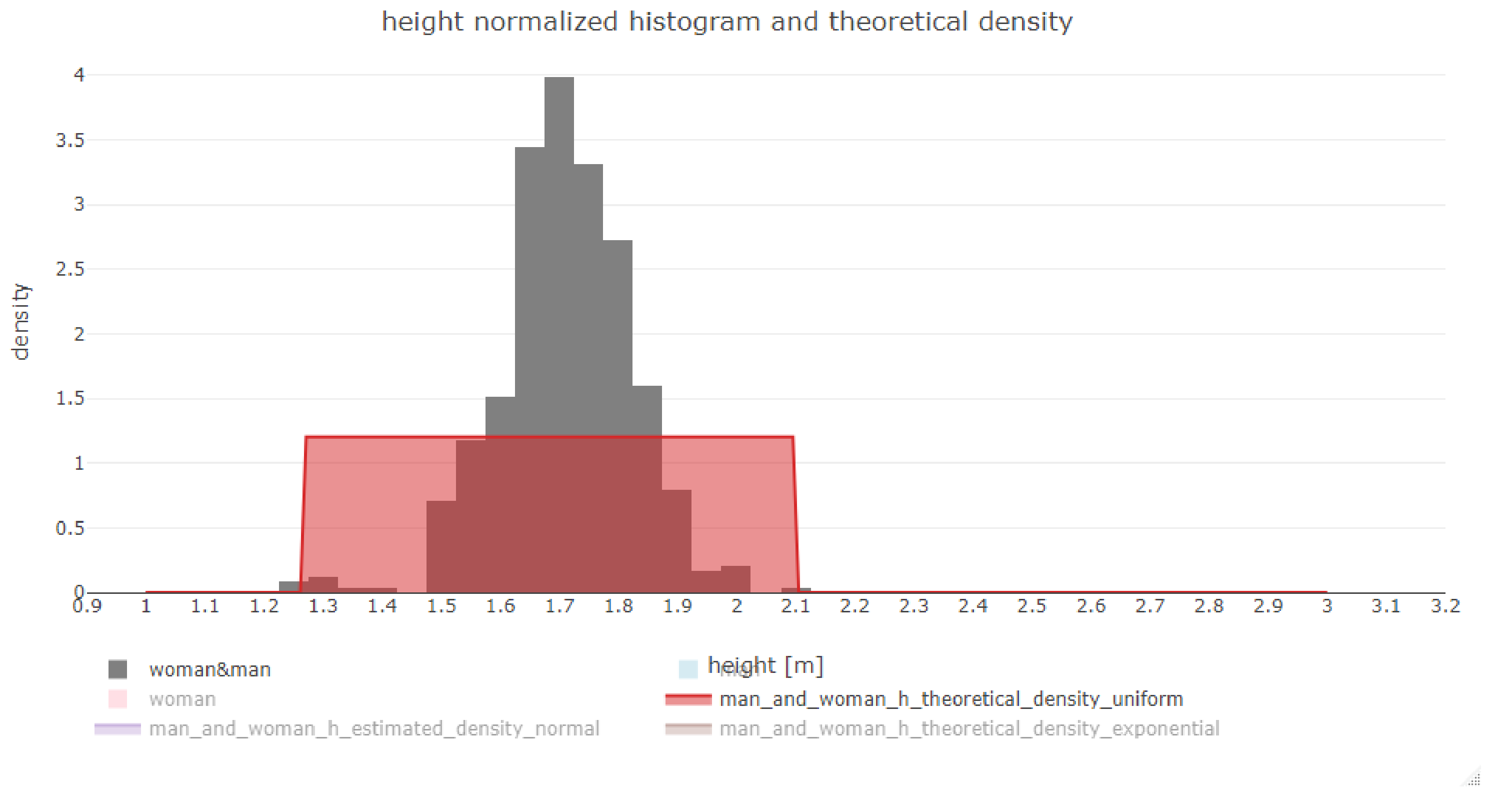
Może rozkład wykładniczy?

```
man_and_woman_h_estimated_theoretical_density_exponential <- EnvStats::eexp(man_and_woman_h)
man_and_woman_h_theoretical_density_exponential <- dexp(x, rate = man_and_woman_h_estimated_theoretical_density_exponential$parameters[1])
```



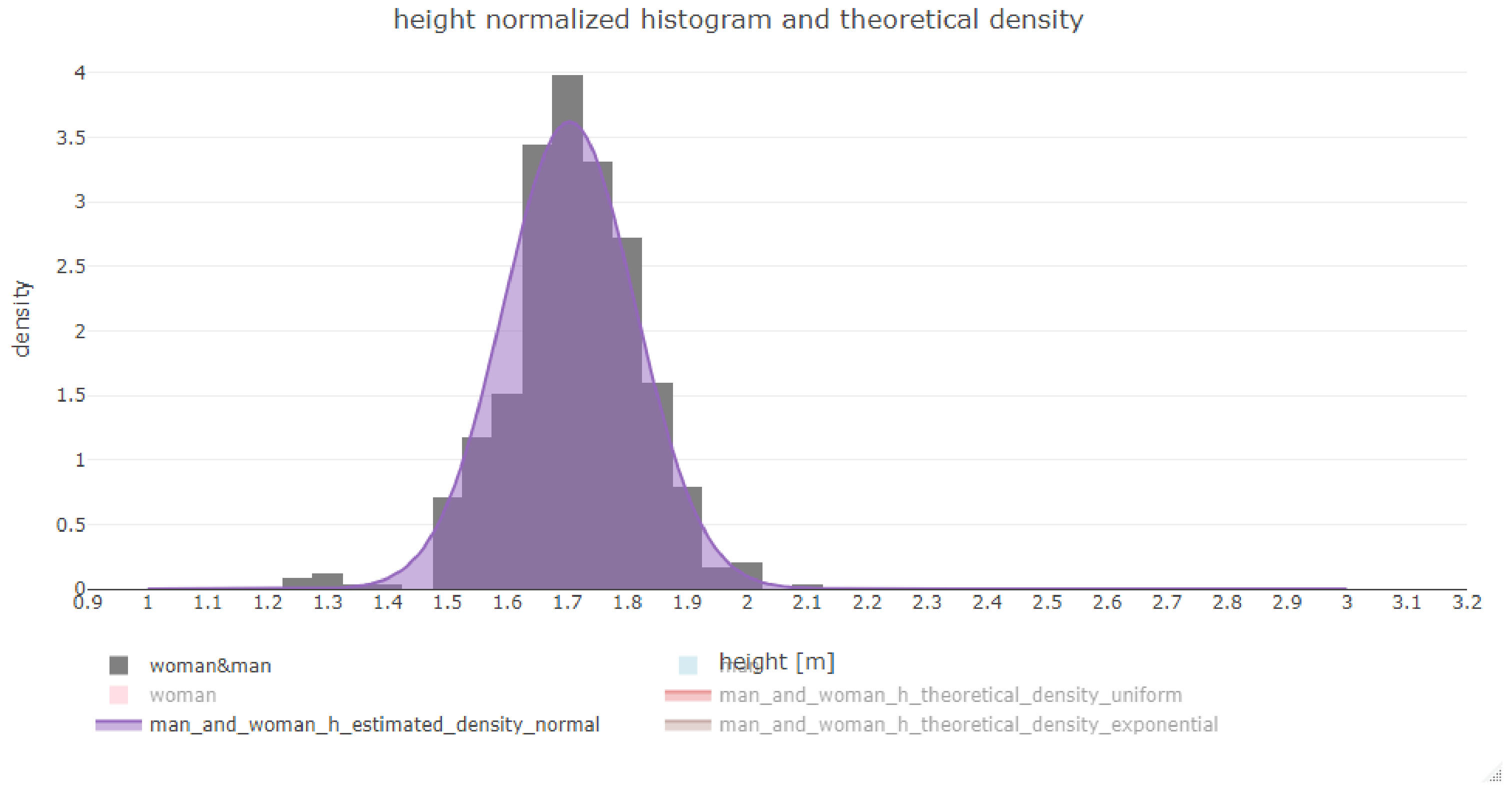
A może rozkład jednostajny?

```
man_and_woman_h_estimated_theoretical_density_uniform <- EnvStats::eunif(man_and_woman_h)
man_and_woman_h_theoretical_density_uniform <- dunif(x,
                                                     min = man_and_woman_h_estimated_theoretical_density_uniform$parameters[1],
                                                     max = man_and_woman_h_estimated_theoretical_density_uniform$parameters[2])
```



A może rozkład normalny?

```
man_and_woman_h_estimated_theoretical_density_normal <- EnvStats::enorm(man_and_woman_h) >
man_and_woman_h_estimated_density_normal <- dnorm(x,
                                          mean = man_and_woman_h_estimated_theoretical_density_normal$parameters[1],
                                          sd = man_and_woman_h_estimated_theoretical_density_normal$parameters[2])
```



Jakie parametry zostały wyestymowane?

```
man_and_woman_h_estimated_theoretical_density_normal$parameters
mean       sd
1.7035743 0.1101773
```

Gęstość wyestymowanego (oszacowanego na podstawie próby) rozkładu teoretycznego (rozkładu dla całej populacji)

$$\boldsymbol{f}_{1.7,0.11}(\boldsymbol{x}) = \frac{\mathbf{1}}{\mathbf{0.11}\sqrt{\mathbf{2}\boldsymbol{\pi}}}\boldsymbol{exp}\left(\frac{-(\boldsymbol{x} - \mathbf{1.7})^{2}}{\mathbf{2} * \mathbf{0.11}^{2}}\right)$$

Oszacujmy teraz jeszcze rozkłady teoretyczne dla kobie i mężczyzn osobno

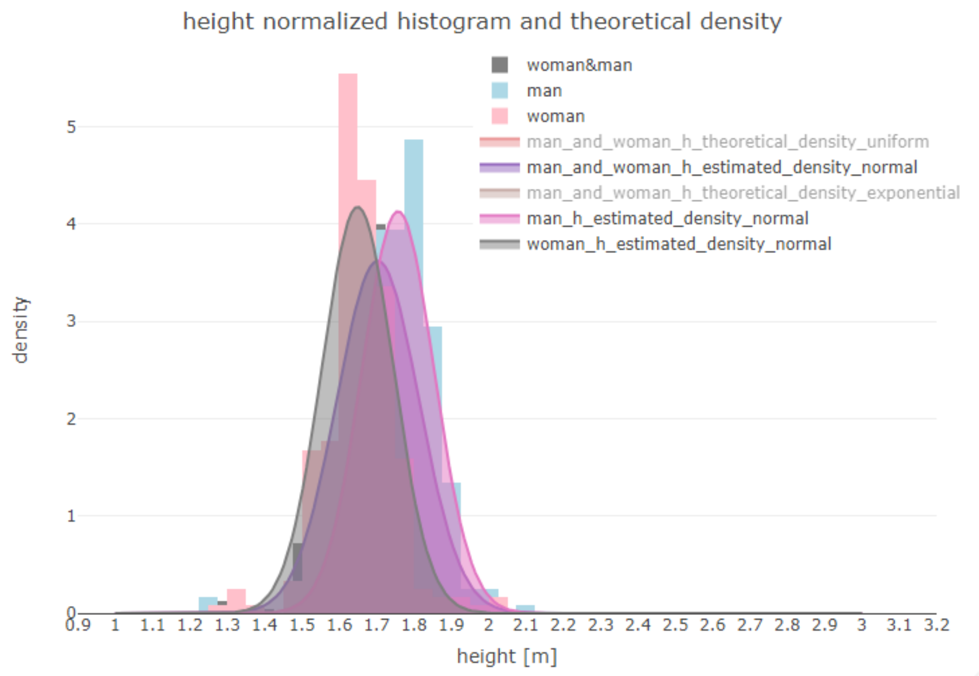
```
man_and_woman_h_estimated_theoretical_density_normal$parameters
mean        sd
1.7035743 0.1101773

man_h_estimated_theoretical_density_normal$parameters
mean        sd
1.75768726 0.09656064

woman_h_estimated_theoretical_density_normal$parameters
mean        sd
1.64946138 0.09552067
```

estymator średniej

estymator odchylenia standardowego



W jaki sposób policzyliśmy estymatory?

Usage

```
enorm(x, method = "mvue", ci = FALSE, ci.type = "two-sided",
  ci.method = "exact", conf.level = 0.95, ci.param = "mean")
```

Arguments

`x`	numeric vector of observations.
`method`	character string specifying the method of estimation. Possible values are `"mvue"` (minimum variance unbiased; the default), and `"mle/mme"` (maximum likelihood/method of moments). See the DETAILS section for more information on these estimation methods.
`ci`	logical scalar indicating whether to compute a confidence interval for the mean or variance. The default value is `FALSE`.
`ci.type`	character string indicating what kind of confidence interval to compute. The possible values are `"two-sided"` (the default), `"lower"`, and `"upper"`. This argument is ignored if `ci=FALSE`.

Generowanie próbek losowych z rozkładu – funkcje rnorm, rexp, ...

Znając oszacowane parametry z rozkładu normalnego, możemy generować **losowe próby** z tego rozkładu (za każdym losowaniem dostaniemy inne wartości, ale wszystkie zadane będą według tego samego rozkładu):

```
rnorm(n = 2, mean = 1.7, sd = 0.11)
[1] 1.844965 1.506314


rnorm(n = 2, mean = 1.7, sd = 0.11)
[1] 1.657440 1.927012
```

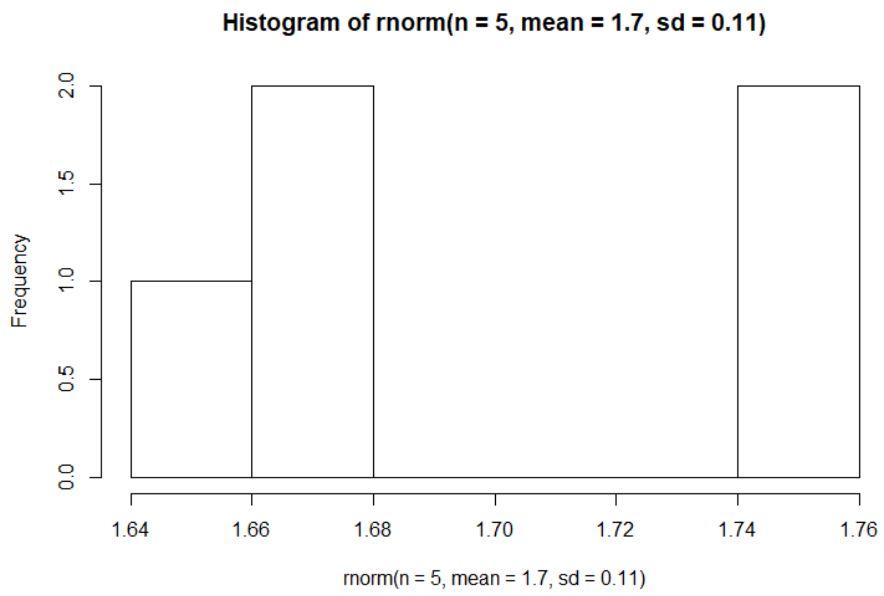
Trick: możemy zapewnić powtarzalność losowania za pomocą „seedów":

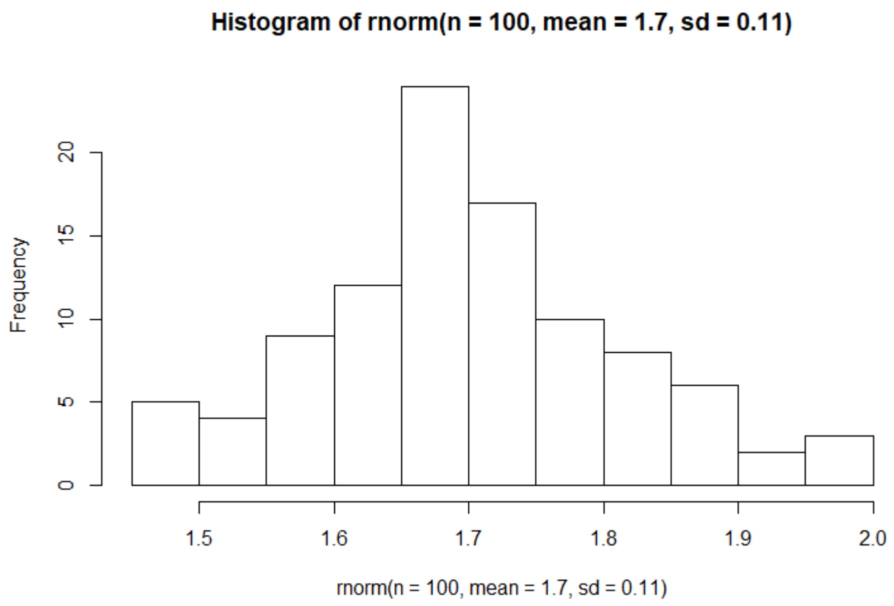
```
set.seed(100)
rnorm(n = 10, mean = 1.7, sd = 0.11)
[1] 1.644759 1.714468 1.691319 1.797546 1.712867 1.735049 1.636003 1.778599 1.609221 [10] 1.660415

set.seed(100)
rnorm(n = 10, mean = 1.7, sd = 0.11)
[1] 1.644759 1.714468 1.691319 1.797546 1.712867 1.735049 1.636003 1.778599 1.609221 [10] 1.660415
```

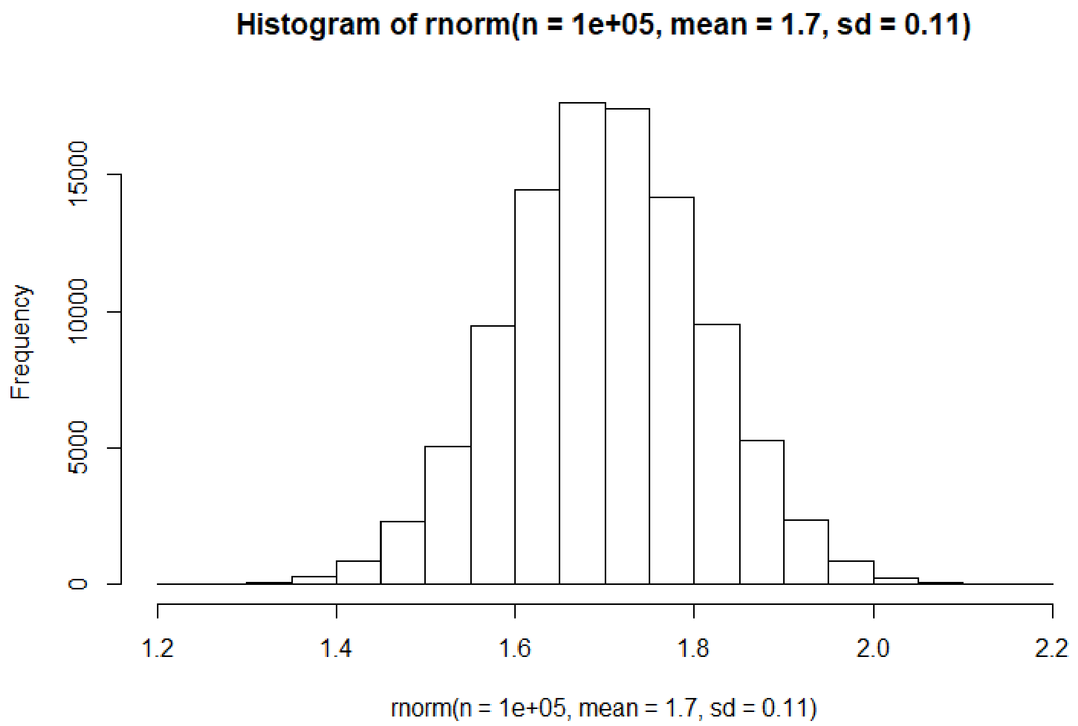
`hist(rnorm(n = 5, mean = 1.7, sd = 0.11))`

`hist(rnorm(n = 100, mean = 1.7, sd = 0.11))`

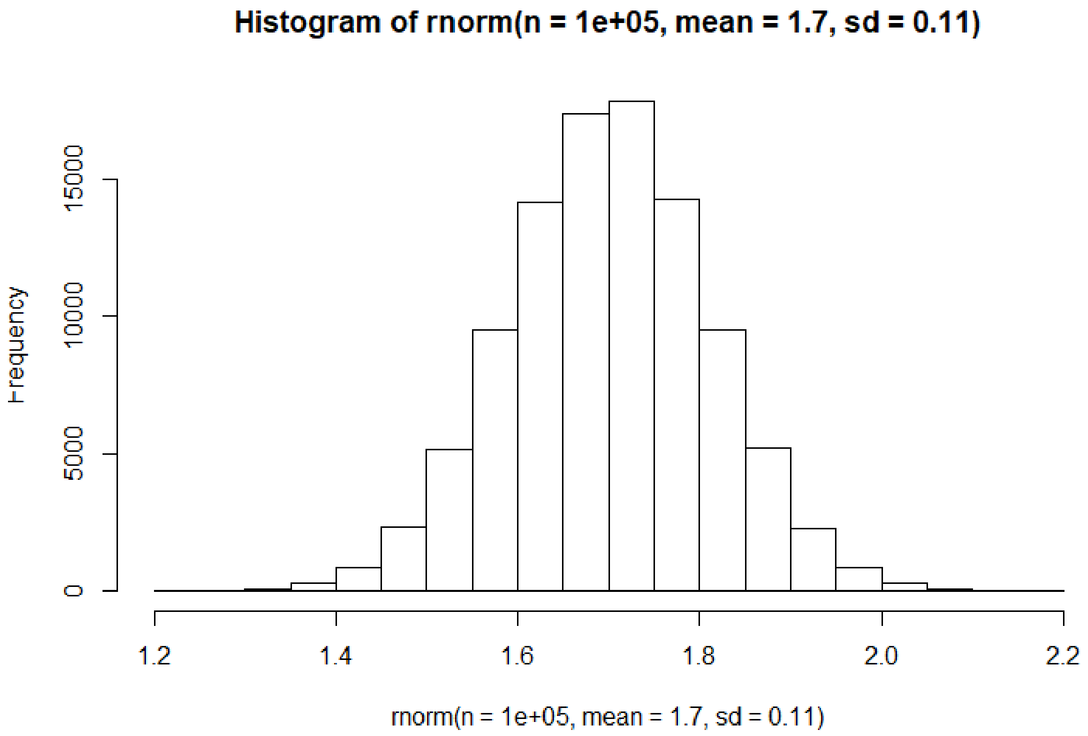




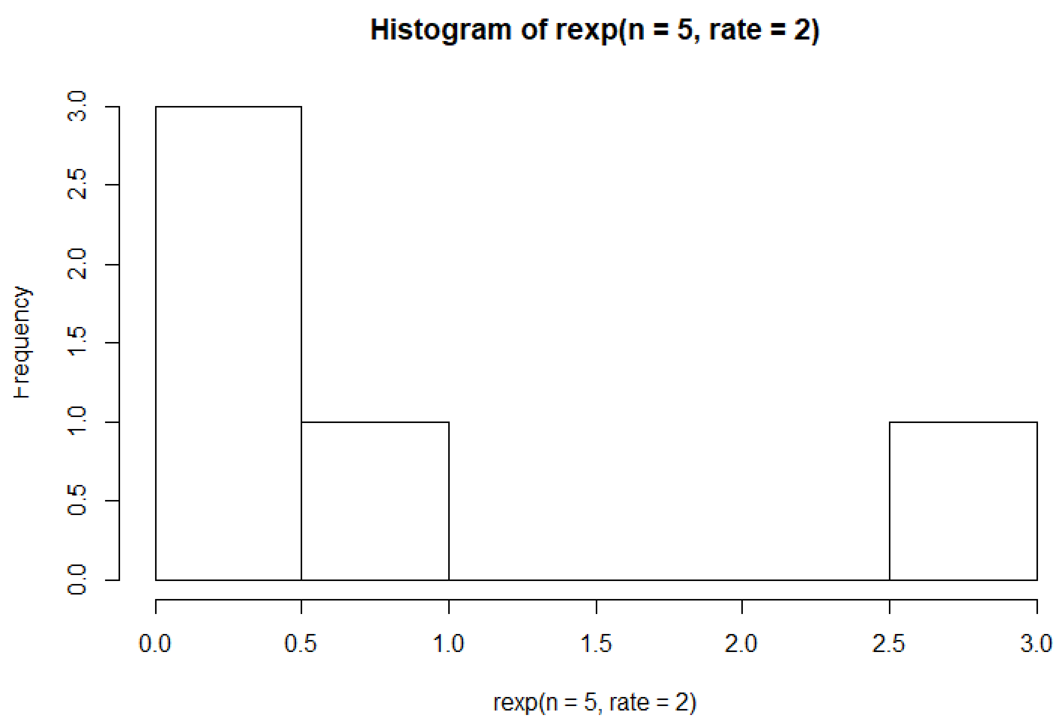
`hist(rnorm(n = 100000, mean = 1.7, sd = 0.11))`



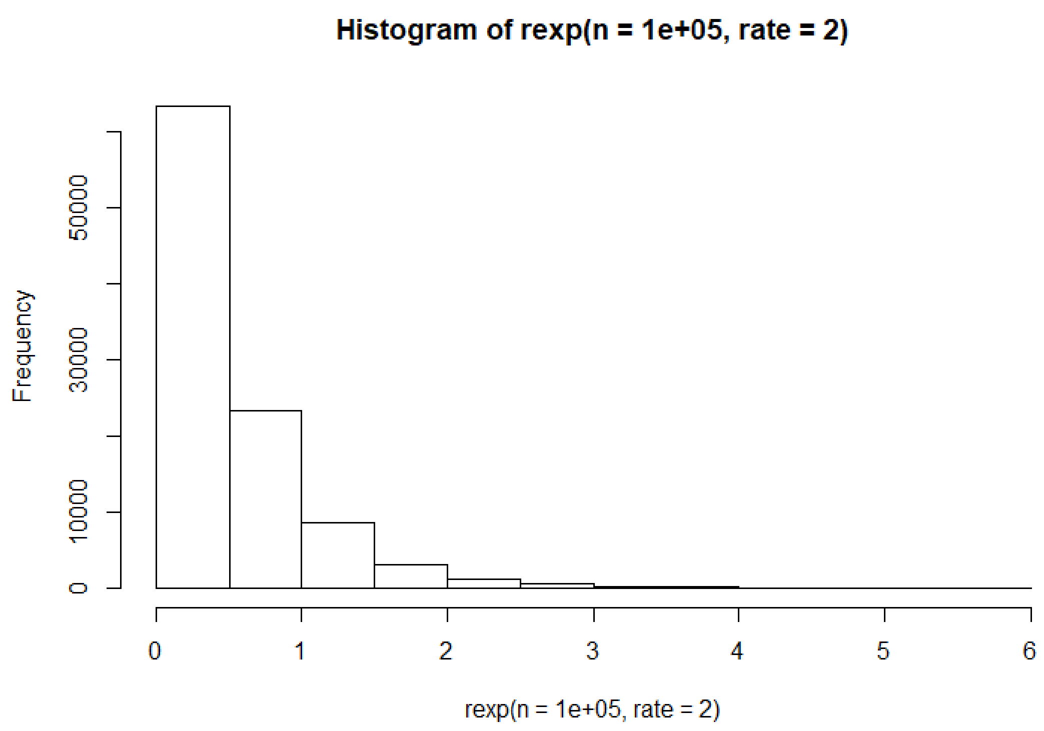
`hist(rnorm(n = 100000, mean = 1.7, sd = 0.11))`



```
hist(rexp(n = 5, rate = 2))
```




```
hist(rexp(n = 100000, rate = 2))
```



THE END:)