ALGORYTMICZNA I STATYSTYCZNA ANALIZA DANYCH

9/10/2014

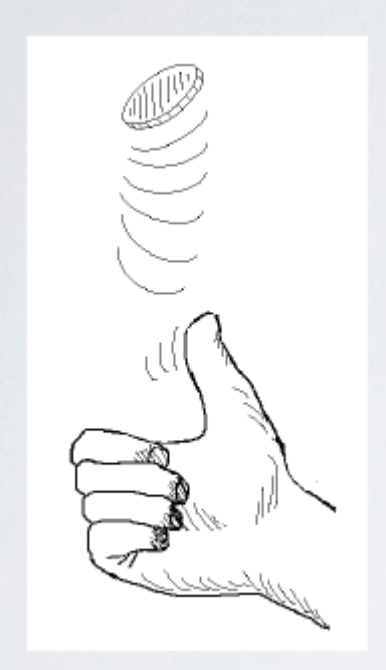
WFAiS UJ, Informatyka Stosowana
II stopień studiów

Prawdopodobieństwo i rozkłady

- Zdarzenia losowe
- Prawdopodobieństwo warunkowe
- Prawdopodobieństwo bayesowskie
- Rozkład normalny
- Rozkład binomialny
- Rozkład binomialny $\rightarrow$ normalny
- Rozkład negatywny binomialny
- Rozkład Poissona

Zdarzenie losowe

- Wiemy jakie zdarzenia mogą nastąpić ale nie wiemy które właśnie nastąpi

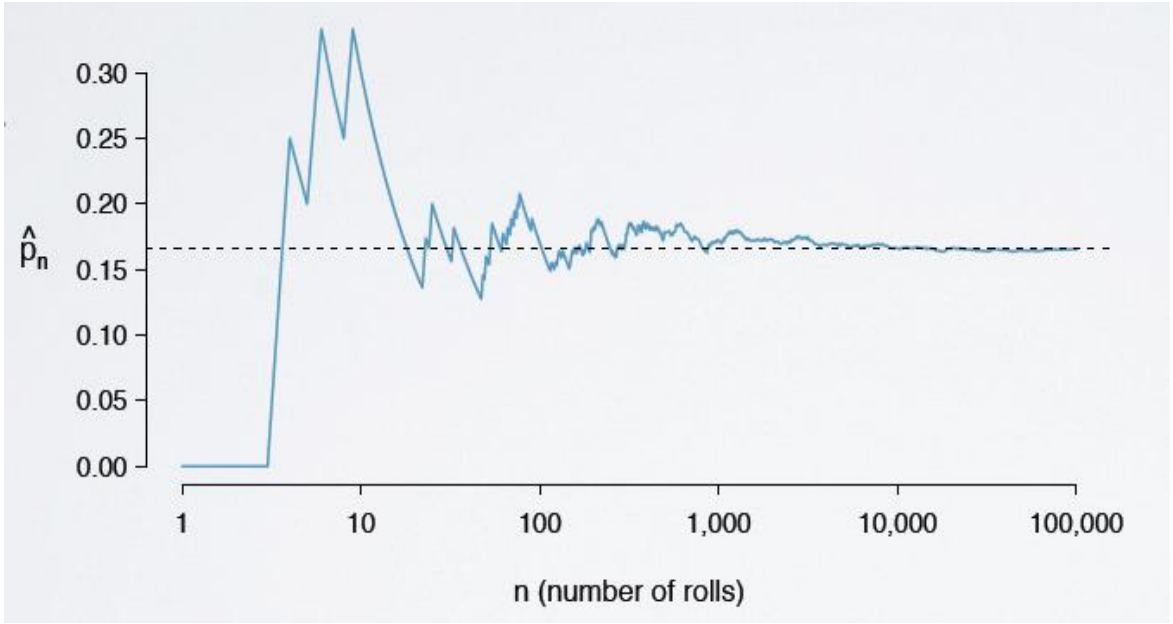


Prawdopodobieństwo

- Jest wiele definicji, ale każda jest zgodna że $0 \leq P(A) \leq 1$
- Definicja „częstościowa” – relatywna ilość zdarzeń w którym obserwujemy dane zdarzenie
- Definicja Bayesowska - stopień naszego zaufania do prawdziwości danej tezy na podstawie informacji która posiadamy

Prawo dużej statystyki

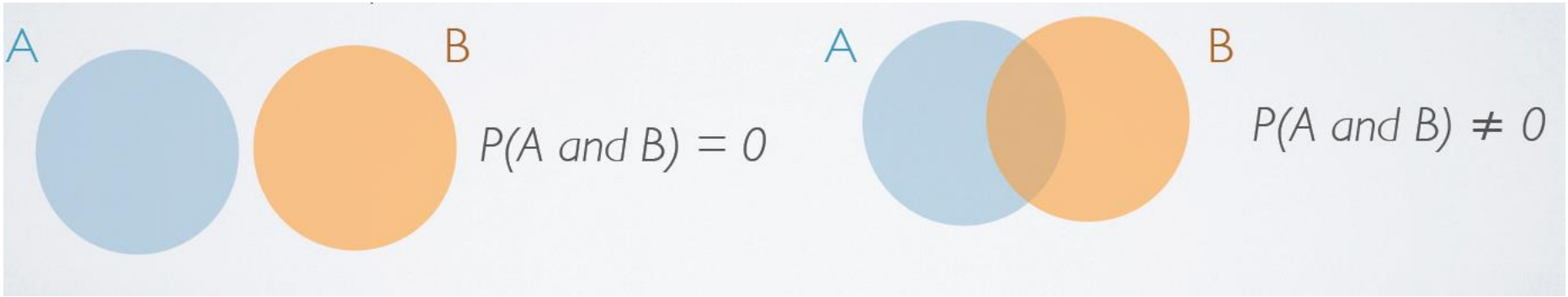
- W miarę jak mamy coraz większą statystykę zdarzeń, proporcja w jakiej pojawia się dane zdarzenie zbliża się do jego prawdopodobieństwa



Proporcja wyrzuconych „1” w funkcji ilości rzutów kostką

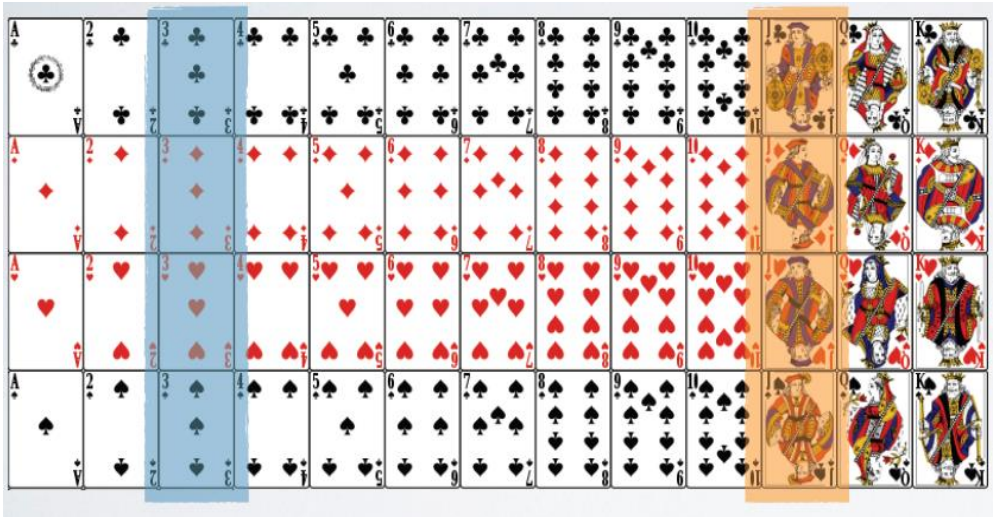
Zdarzenia rozłączne i nie-rozłączne

- Rozłączne (wzajemnie wykluczające się) zdarzenia nie mogą zajść w tym samym czasie
  - Możemy wyrzucić albo reszkę albo orzełka
  - Student nie może równocześnie zdać i oblać egzaminu



Suma rozłącznych zdarzeń

- Jakie jest prawdopodobieństwo aby wyciągnąć „Jokera” lub „3” z tali kart



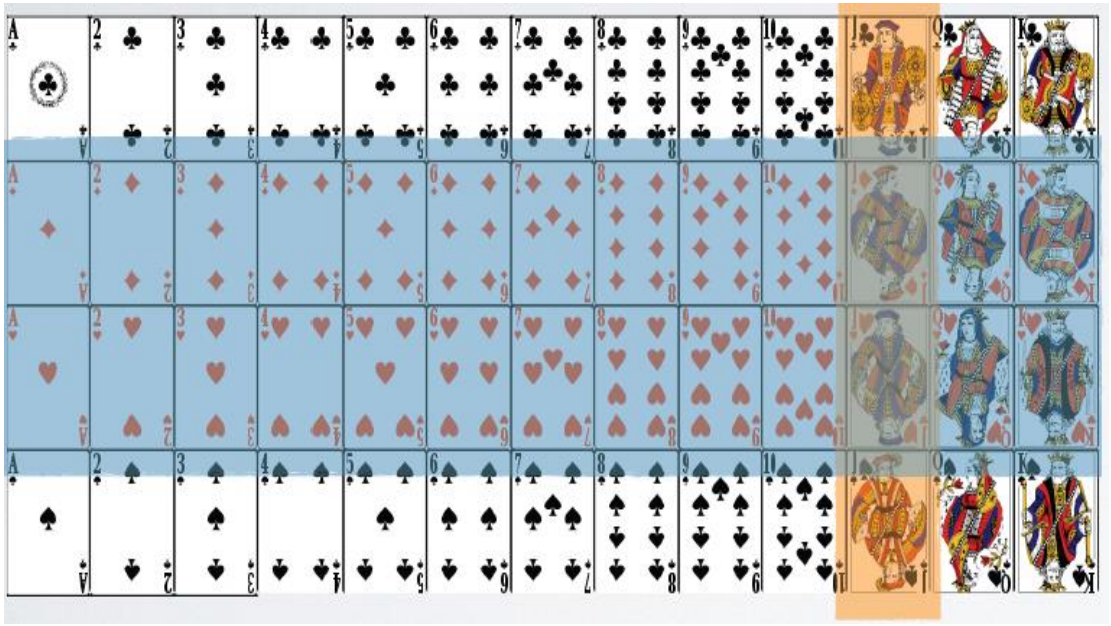
$$P(J \text{ or } 3)$$
$$= P(J) + P(3)$$
$$= (4/52) + (4/52)$$
$$\approx 0.154$$

Dla rozłącznych zdarzeń
P(A or B) = P(A) + P(B)

Suma nie-rozłącznych zdarzeń

- Jakie jest prawdopodobieństwo aby wyciągnąć „Jokera” lub „czerwonej karty”

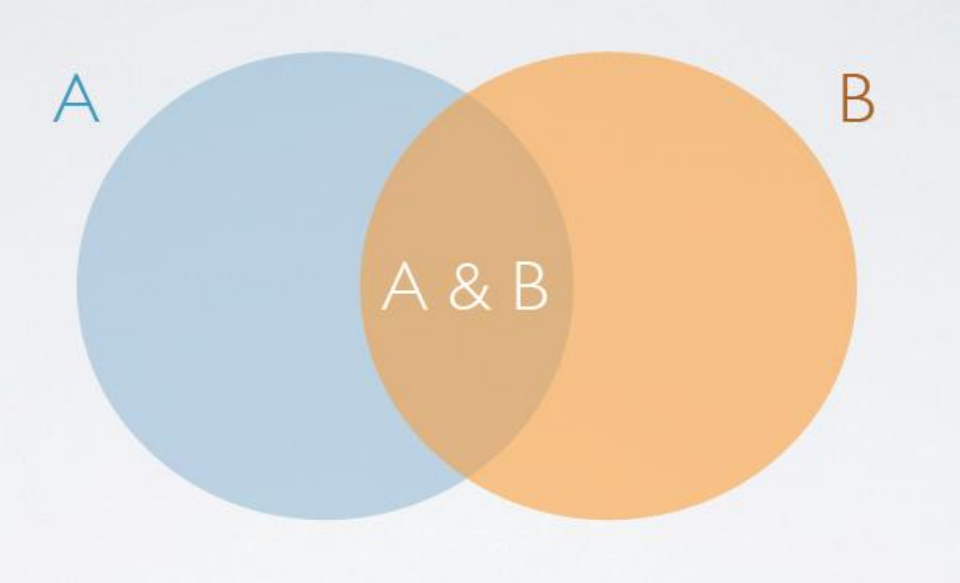
  karty z tali kart



$$P(J \text{ or red})$$
$$= P(J) + P(\text{red}) - P(J \text{ and red})$$
$$= (4/52) + (26/52) - (2/52)$$
$$\approx 0.538$$

Dla nie-rozłącznych zdarzeń
P(A or B) = P(A) + P(B)
– P(A and B)

Suma zdarzeń



P(A or B) = P(A) + P(B) – P(A and B)

Przestrzeń zdarzeń

- Zbiór wszystkich możliwych zdarzeń
- Np. mamy dwoje dzieci, jakiej mogą być płci?

S = { MM, KK, MK, KM}

Przestrzeń zdarzeń

Rozkład prawdopodobieństwa

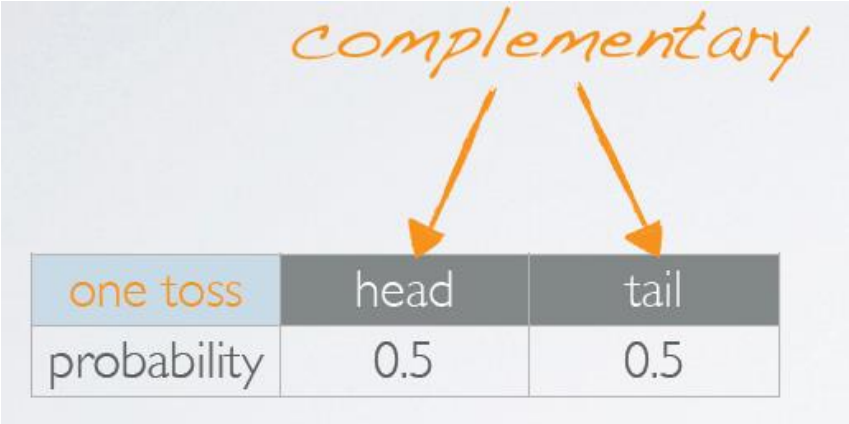
one toss	head	tail
probability	0.5	0.5

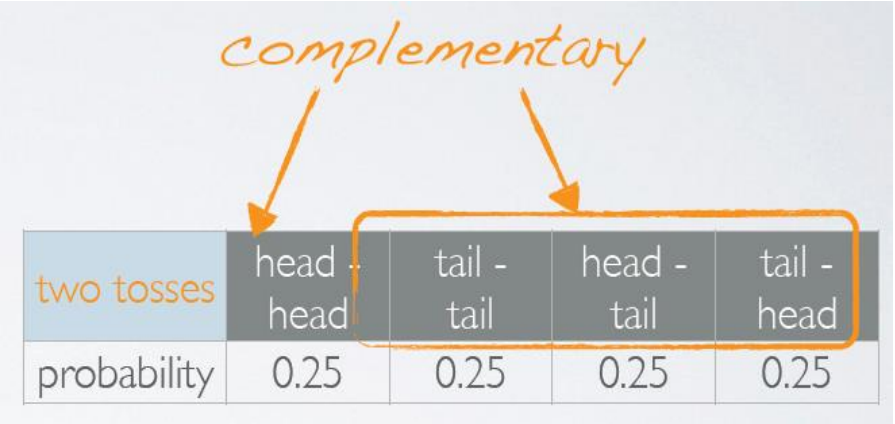
two tosses	head - head	tail - tail	head - tail	tail - head
probability	0.25	0.25	0.25	0.25

- Lista prawdopodobieństw wszystkich możliwych wyników w przestrzeni zdarzeń
- Reguły:
  - Zdarzenia muszą być rozłączne
  - Każde z prawdopodobieństw pomiędzy 0 i 1
  - Suma prawdopodobieństw = 1

Zdarzenia uzupełniające się

To są dwa wzajemnie wykluczające się zdarzenia których suma prawdopodobieństw = 1





Rozłączne vs uzupełniające się

Suma dwóch zdarzeń uzupełniających się jest zawsze równa 1.
Ale to nie musi być prawdą dla dwóch zdarzeń rozłącznych.

Zdarzenia zależne i niezależne

Dwa zdarzenie są niezależne jeżeli wynik jednego nie daje żadnej informacji na temat wyniku drugiego

niezależne

zależne

Kiedy są niezależne

Sprawdzamy: jeżeli $P(A \mid B) = P(A)$ to zdarzenia są niezależne

Obserwujemy różnicę w prawdopodobieństwach warunkowych → zależność → testujemy taką hipotezę

Jeżeli różnica jest duża
To jest to silna wskazówka ze zdarzenia są zależne

Przy dużej statystyce nawet mała różnica może wskazywać na zależność

Przykład

Jak młodzi ludzie postrzegają przynależność do warstw społecznych

results:		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle class	8	37	45
	upper class	0	0	0
	Total	48	50	98

Study reference: Goodman, Elizabeth, et al. "Adolescents' understanding of social class: a comparison of white upper middle class and working class youth." Journal of adolescent health 27.2 (2000): 80-83.

Przykład: globalna charakterystyka

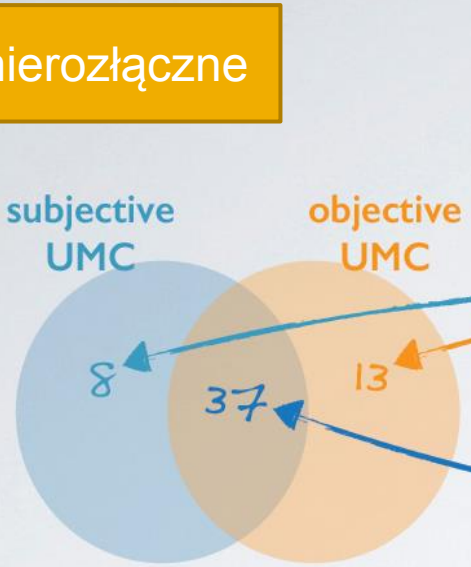
		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle class	8	37	45
	upper class	0	0	0
	Total	48	50	98

Prawdopodobieństwo że uczeń obiektywnie należy do UMC

$P(obj\ UMC) = 50 / 98 \approx 0.51$

Przykład: zdarzenia nierozłączne

nierozłączne



		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle	8	37	45
	upper class	0	0	0
	Total	48	50	98

Prawdopodobieństwo że obiektywnie i subiektywnie uczeń należy do UMC

$P(\text{obj UMC} \,\&\, \text{subj UMC}) = 37 / 98 \approx 0.38$

Przykład: zdarzenia warunkowe

		social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle	8	37	45
	upper class	0	0	0
	Total	48	50	98

Prawdopodobieństwo że uczeń który obiektywnie jest w WC subiektywnie należy do UMC

$P(\text{subj UMC} \mid \text{obj WC}) = 8 / 48 \approx 0.17$

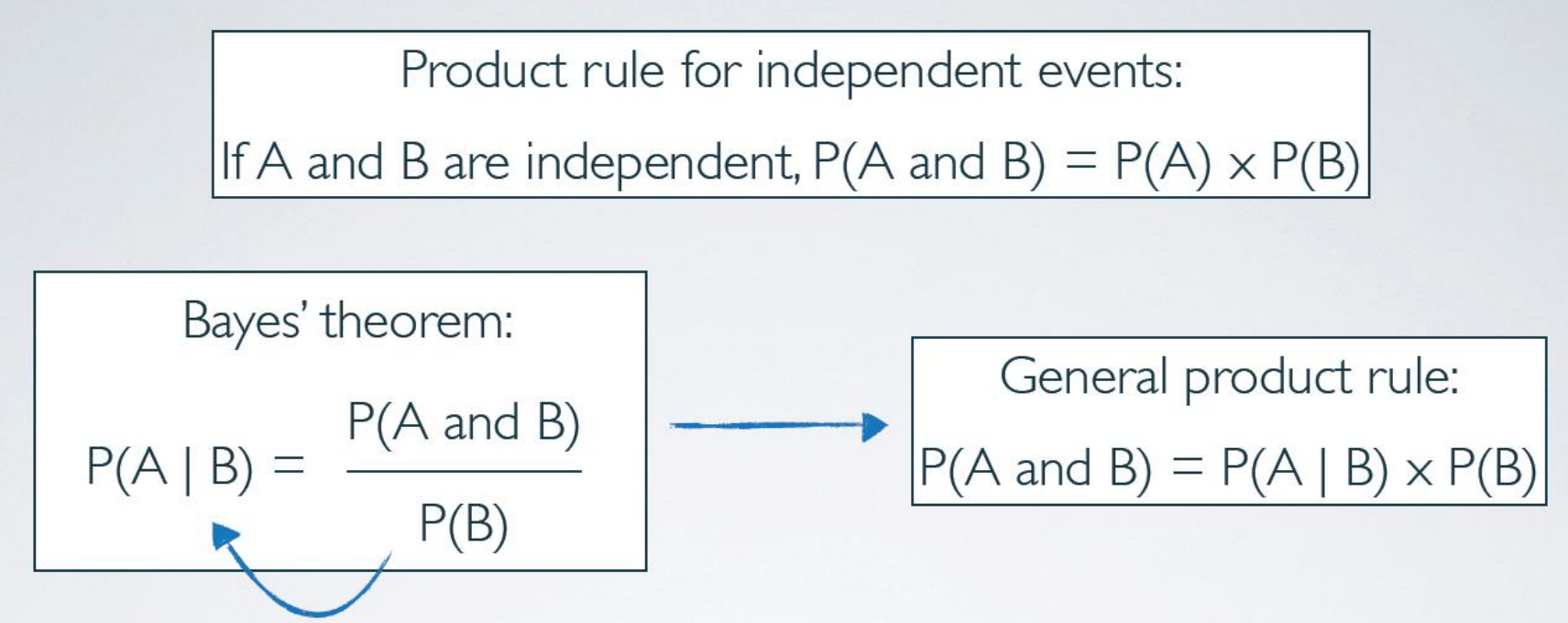
Prawdopodobnie Bayesowskie

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle	8	37	45
	upper class	0	0	0
	Total	48	50	98

$$P(\text{subj UMC} \mid \text{obj WC}) = \frac{P(\text{subj UMC \& obj WC})}{P(\text{obj WC})} = \frac{8/98}{48/98} = 8/48 \approx 0.17$$

Niezależne zdarzenia



Niezależność a prawdopodobieństwo warunkowe

Generalnie jeżeli P(A | B) = P(A) to zdarzenia są niezależne

- Intuicyjnie:
  - Znajomość B nic nam nie mówi o A
- Matematycznie:
  - jeżeli A i B są niezależne to

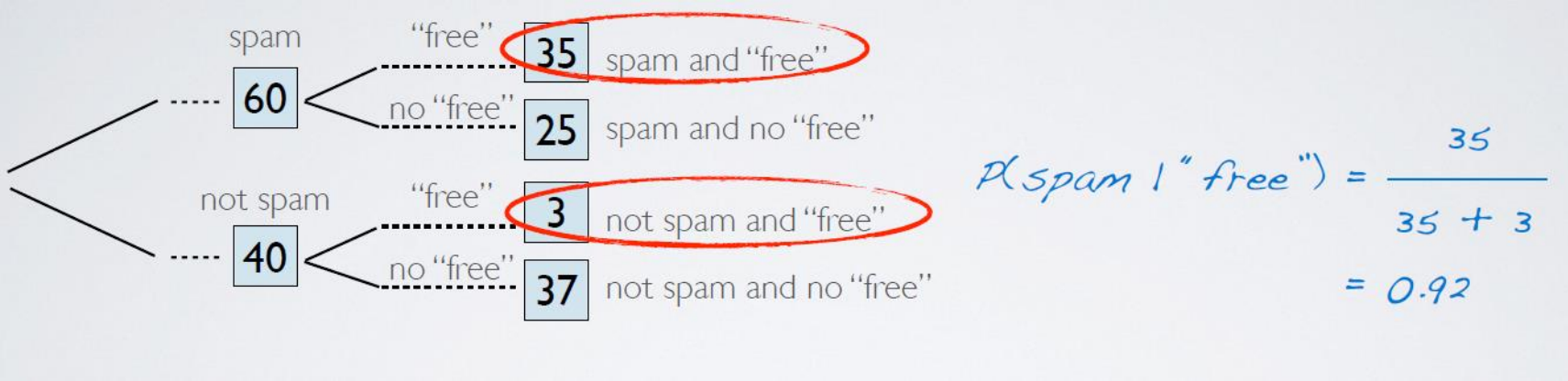
$$P(A \text{ and } B) = P(A) \times P(B),$$

wtedy

$$P(A \mid B) = P(A \text{ and } B)/P(B) = P(A)$$

Drzewo prawdopodobieństw

Mamy 100 emaili: 60 to „spam”, 40 to „not spam”.
Z 60 spam emaili, 35 zawiera słowo „free”, z pozostałych tylko 3 zawierają słowo „free”.
Pytanie: Jeżeli email zawiera słowo „free” jakie jest prawdopodobieństwo że jest to spam?



Wnioskowanie Bayesowskie

- Przykład:

Jakie jest prawdopodobieństwo że wyrzucimy ≥ 4 ?

Wnioskowanie Bayesowskie

Jakie jest prawdopodobieństwo że wyrzucimy ≥ 4 ?

$$S = \{1,2,3,4,5,6\}$$

$$P(\geq 4) = 3/6 = 1/2 = 0.5$$

$$S = \{1,2,3,4,5,6,7,8,9,10,11,12\}$$

$$P(\geq 4) = 9/12 = 3/4 = 0.75$$

Wnioskowanie Bayesowskie

- Reguły

Wnioskowanie Bayesowskie

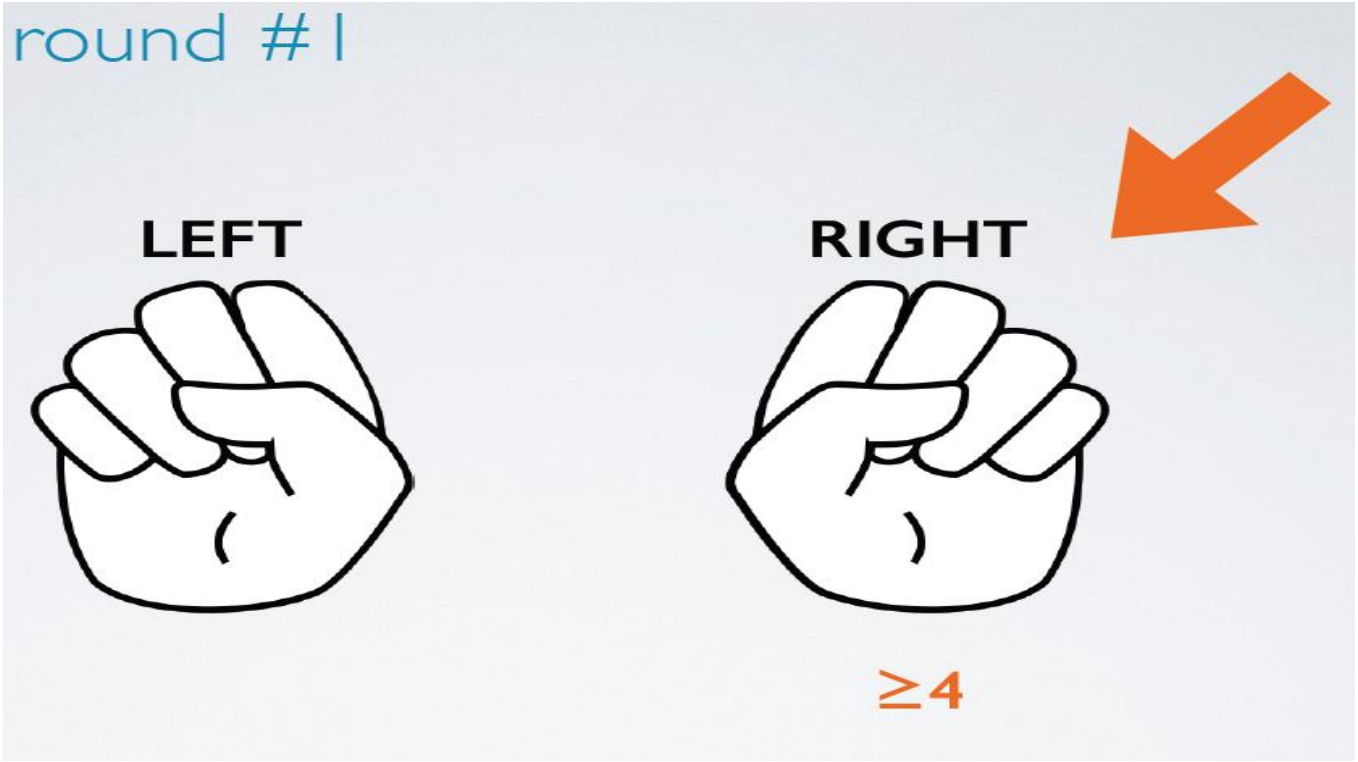
- Zanim zaczniemy zbierać dane jakie są prawdopodobieństwa dla następujących hipotez:
  - H_1 – „dobra” kostka w lewej ręce
  - H_2 - „dobra” kostka w prawej ręce

	P(H_1: good die on the Right)	P(H_2: good die on the Left)
(a)	0.33	0.67
(b)	0.5	0.5
(c)	0	1
(d)	0.25	0.75

→ prior

Wnioskowanie Bayesowskie

- Zbieramy dane



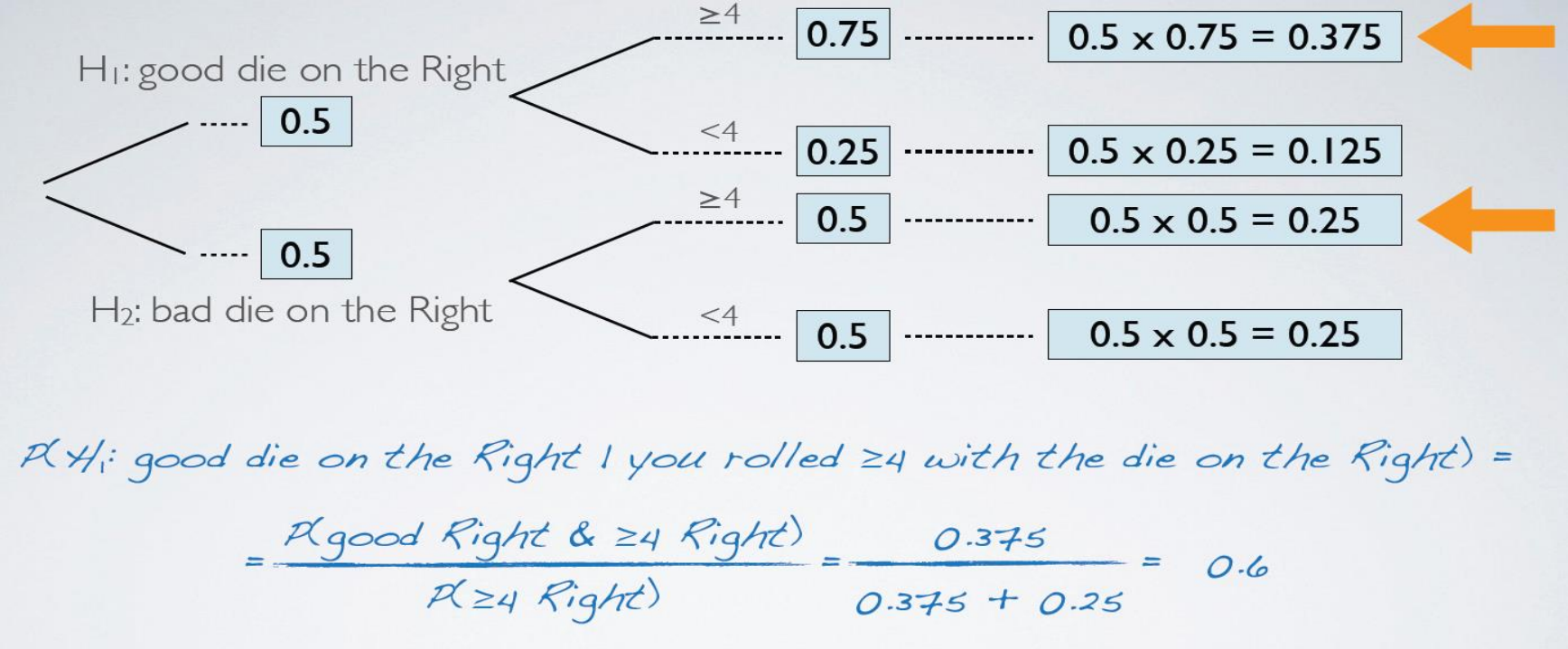
Wnioskowanie Bayesowskie

- Po tym jak zobaczyliśmy dane możemy zmienić prawdopodobieństwa

	$P(H_1$: good die on the Right)	$P(H_2$: good die on the Left)
(a)	0.5	0.5
(b)	more than 0.5	less than 0.5
(c)	less than 0.5	more than 0.5

Wnioskowanie Bayesowskie

- Możemy to też policzyć wiedząc jaki był wynik pierwszej rundy



Wnioskowanie Bayesowskie

- Prawdopodobieństwo które wyliczyliśmy jest nazywane „posterior” prawdopodobieństwem
- Jest ono oznaczane też P(hipoteza| dana), czyli prawdopodobieństwo dla hipotezy biorąc pod uwagę dostępne już dane.
- Zależy od założonego prawdopodobieństwa „prior” i zebranych danych.
- Do kolejnej rundy poprawiamy nasz „prior” wykorzystując znany już posterior

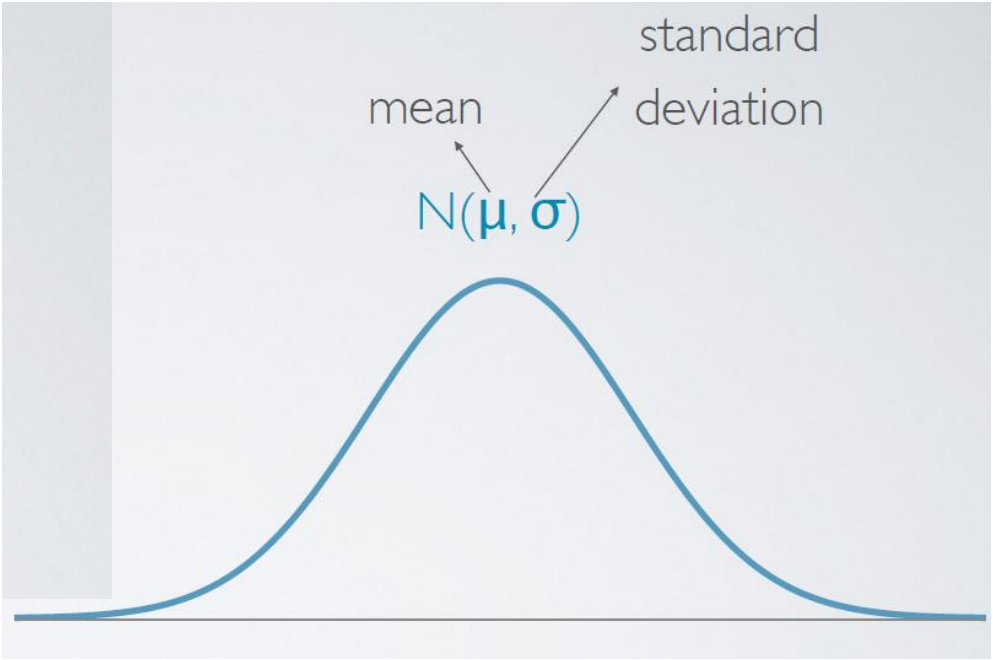
updated:	$P(H_1$: good die on the Right)	$P(H_2$: good die on the Left)
	0.6	0.4

Wnioskowanie Bayesowskie

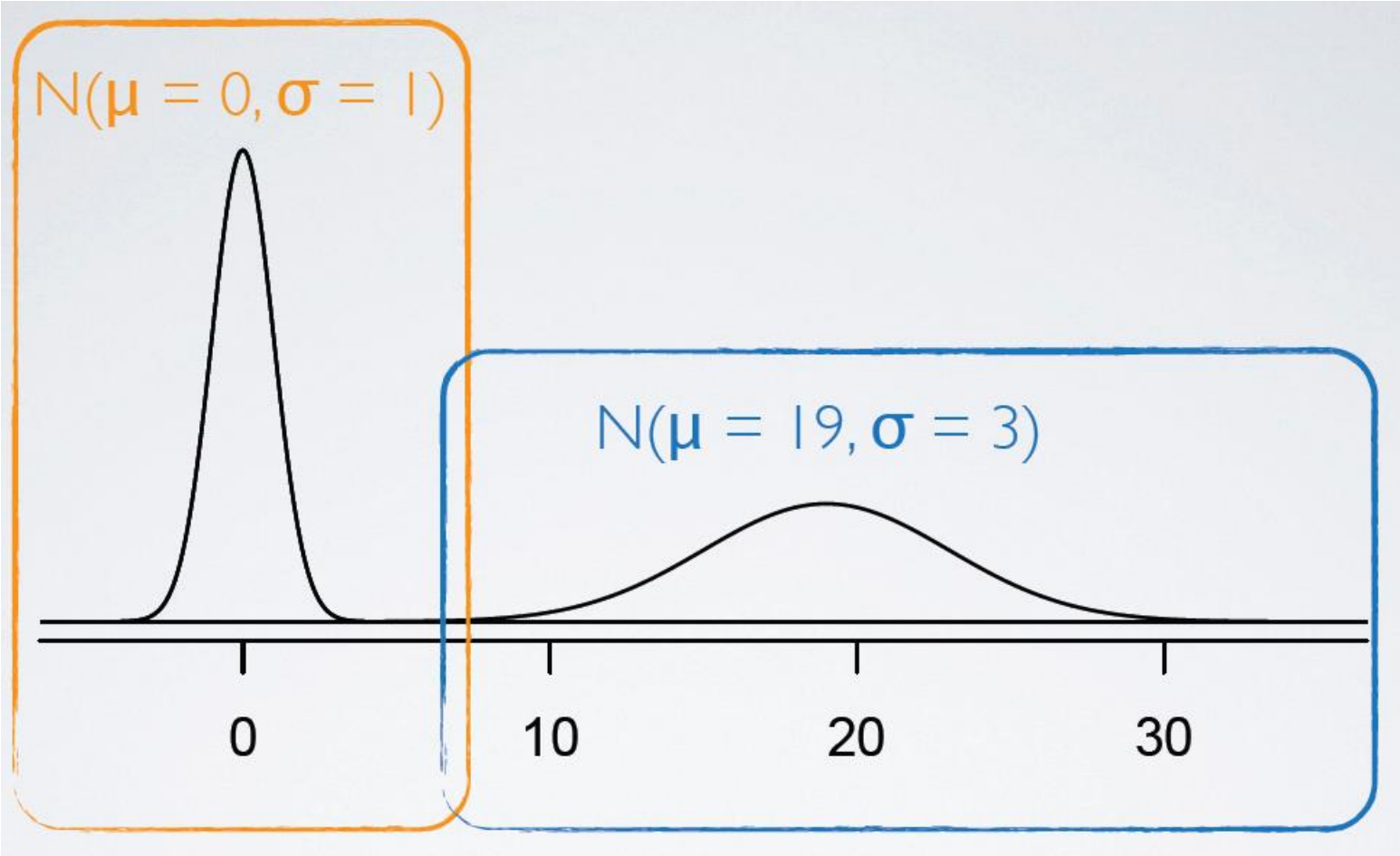
- Wykorzystuje znaną już informację, jak np. model lub wykonane już badania
- W naturalny sposób włącza informacje z danych w trakcie ich zbierania
- Opiera decyzje na „posterior” prawdopodobieństwie:
  - P(H jest prawdziwa | zebrane dane)
- Dobry początkowy „prior” pomaga, „zły” przeszkadza coraz mniej w miarę jak zbieramy dane
- Bardziej zaawansowane techniki bayesowskie stwarzają szersze możliwości niż tradycyjne podejście „częstościowe”.

Rozkład normalny

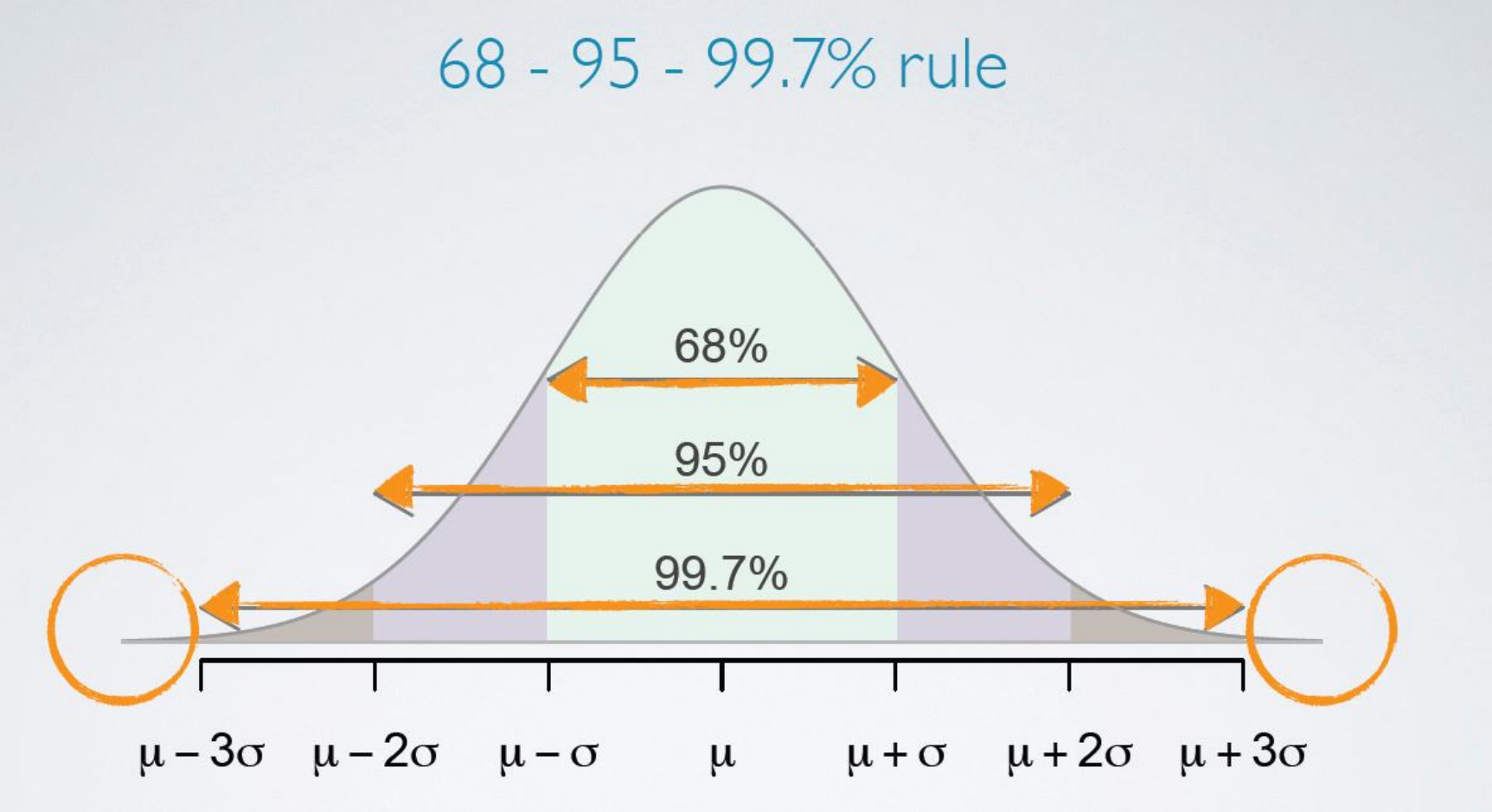
- Symetryczny
- Jednomodalny
- Dokładna reguła jaka jest zmienność danych wokół wartości średniej
- Wiele zmiennych ma rozkład zbliżony do normalnego.



Rozkład normalny



Rozkład normalny



Praktycznie jak wykorzystujemy

- Lekarz podaje wyniki pomiaru pulsu przebadanej próbki pacjentów która ma rozkład zbliżony do normalnego. Podaje tylko trzy liczby: średnia = 110/min, minimalna = 65/min, maksymalna = 155/min.
- Jakie jest najbardziej prawdopodobnie odchylenie standardowe?

(a) 5 → $110 \pm (3 \times 5) = (95, 125)$

(b) 15 → $110 \pm (3 \times 15) = (65, 155)$

(c) 35 → $110 \pm (3 \times 35) = (5, 215)$

(d) 90 → $110 \pm (3 \times 90) = (-160, 380)$

Porównywanie rozkładów

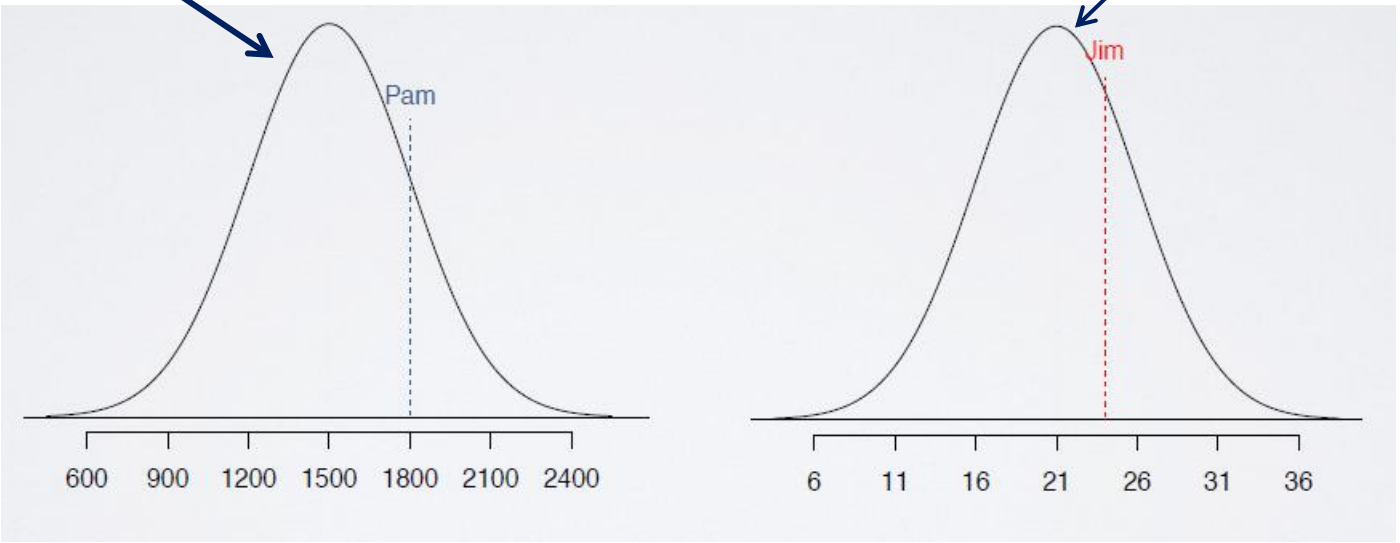
Ilość punktów zebranych przez studentów w różnych konkursach.
Czy możemy ich porównywać ze sobą na podstawie tych wyników?

średnia

odchylenie standardowe

$$\text{SAT scores} \sim N(\text{mean} = 1500, SD = 300)$$

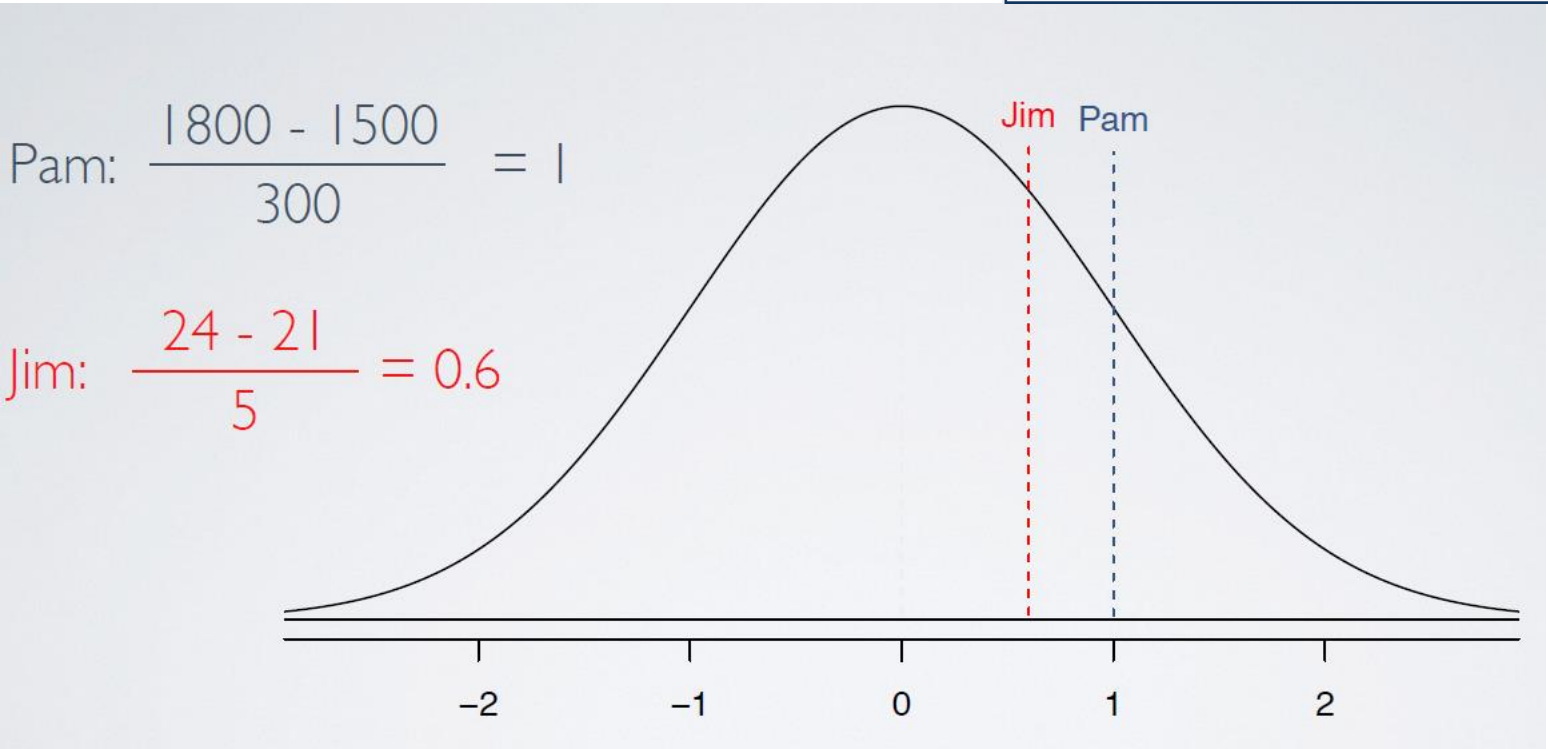
$$\text{ACT scores} \sim N(\text{mean} = 21, SD = 5)$$



Porównywanie rozkładów

Standaryzujemy rozkłady:

$$Z = \frac{\text{obserwacja - średnia}}{\text{odchylenie standardowe}}$$

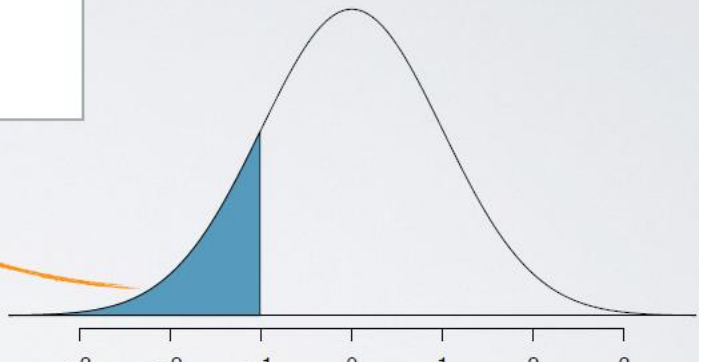


Percentile

Dla standaryzowanego rozkładu możemy wyliczyć percentile, czyli procent obserwacji które są mniejsze niż dana wartość.

Korzystamy z funkcji języka R

```
R
> pnorm(-1, mean = 0, sd = 1)
[1] 0.1586553
```

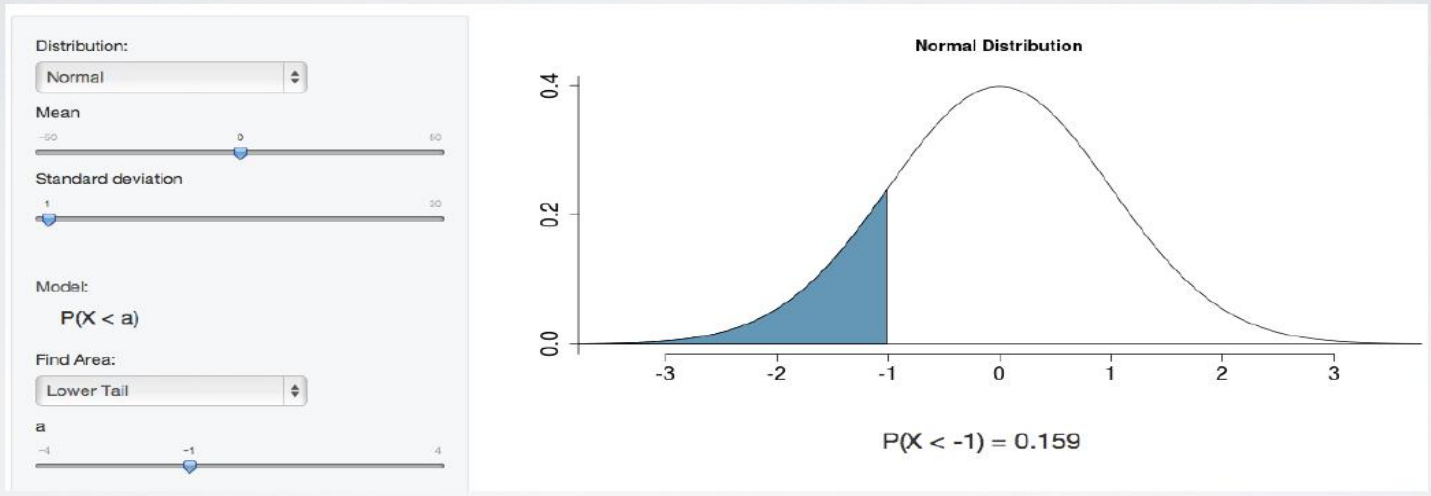


Percentile

Dla standaryzowanego rozkładu możemy wyliczyć percentile, czyli procent obserwacji które są mniejsze niż dana wartość.

Korzystamy z appletu

http://bitly.com/dist_calc

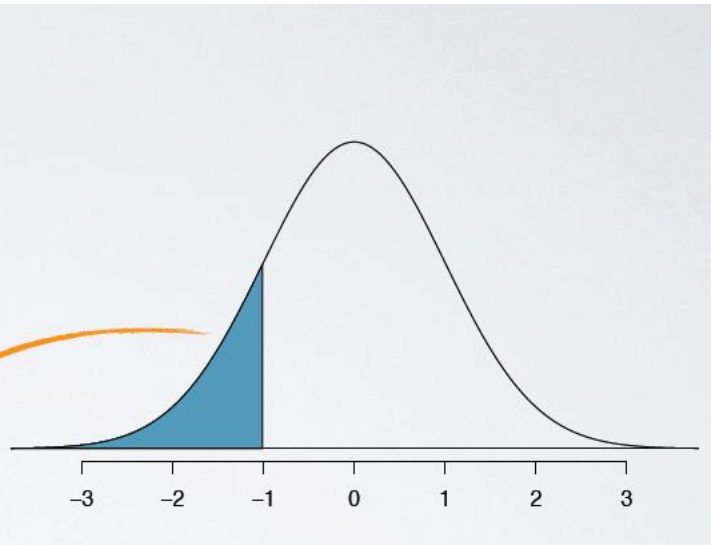


Percentile

Dla standaryzowanego rozkładu możemy wyliczyć percentile, czyli procent obserwacji które są mniejsze niż dana wartość.

Korzystamy z tabeli

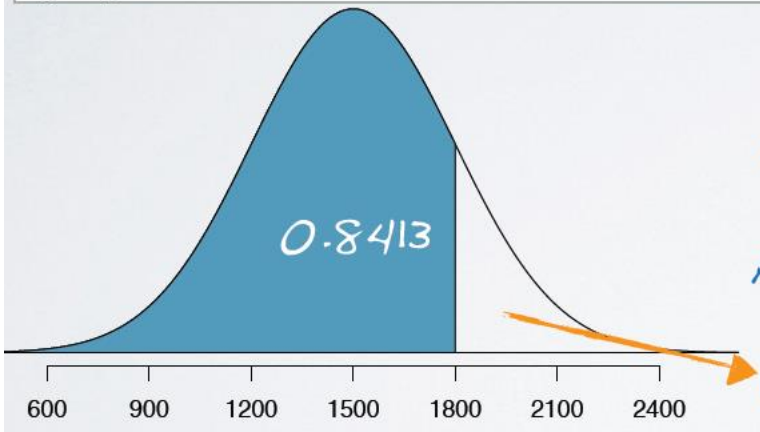
Second decimal place of Z					
0.04	0.03	0.02	0.01	0.00	Z
0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0618	0.0630	0.0643	0.0655	0.0668	−1.5
0.0749	0.0764	0.0778	0.0793	0.0808	−1.4
0.0901	0.0918	0.0934	0.0951	0.0968	−1.3
0.1075	0.1093	0.1112	0.1131	0.1151	−1.2
0.1271	0.1292	0.1314	0.1335	0.1357	−1.1
0.1492	0.1515	0.1539	0.1562	0.1587	−1.0



Percentile

Jaki procent uczniów otrzymał większa ilość punktów niż Pam

```
R
> pnorm(1800, mean = 1500, sd = 300)
[1] 0.8413
```



$$Z = \frac{1800 - 1500}{300} = 1$$

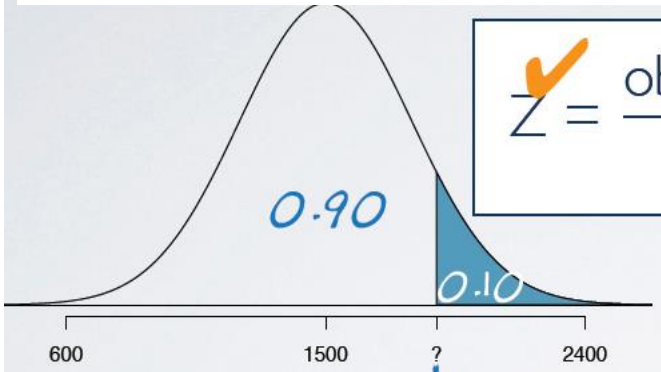
$$P(Z < 1) = 0.8413$$

$$1 - 0.8413 = 0.1587$$

	Second decimal place of Z			
Z	0.00	0.01	0.02	
0.0	0.5000	0.5040	0.5080	0.
0.1	0.5398	0.5438	0.5478	0.
0.2	0.5793	0.5832	0.5871	0.
0.8	0.7881	0.7910	0.7939	0.
0.9	0.8159	0.8186	0.8212	0.
1.0	0.8413	0.8438	0.8461	0.
1.1	0.8643	0.8665	0.8686	0.

Percentile

A jeżeli ktoś mówi że jest top 10% w SAT klasyfikacji. To ile otrzymał punktów ?



$$Z = \frac{\text{observation - mean}}{\text{SD}}$$

	Second decimal place of Z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
			0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
			0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
			0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
			0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
			0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
			0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

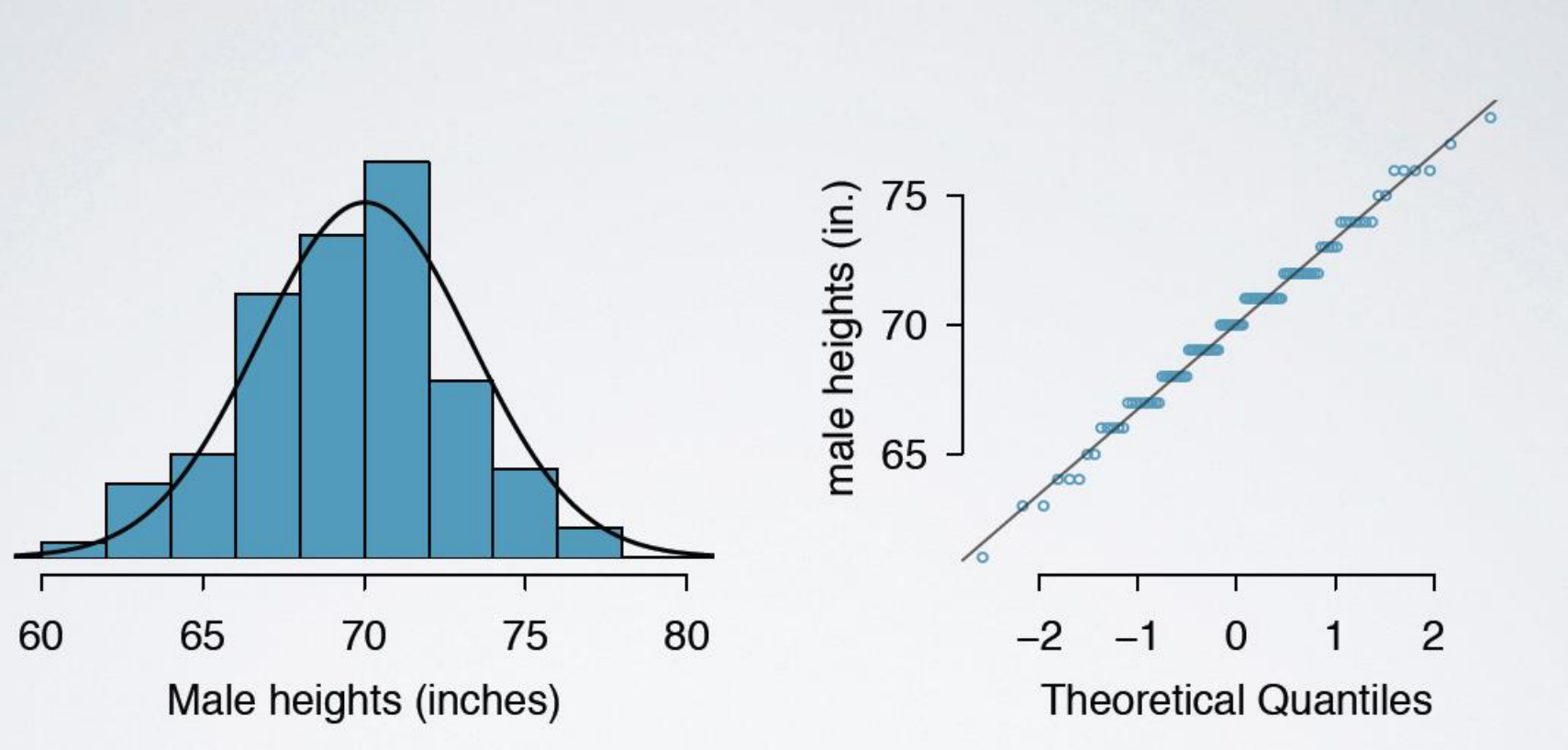
$$Z = 1.28 = \frac{X - 1500}{300}$$

$$X = (1.28 \times 300) + 1500 = 1884$$

R

```
> qnorm(0.90, 1500, 300)
[1] 1884.465
```

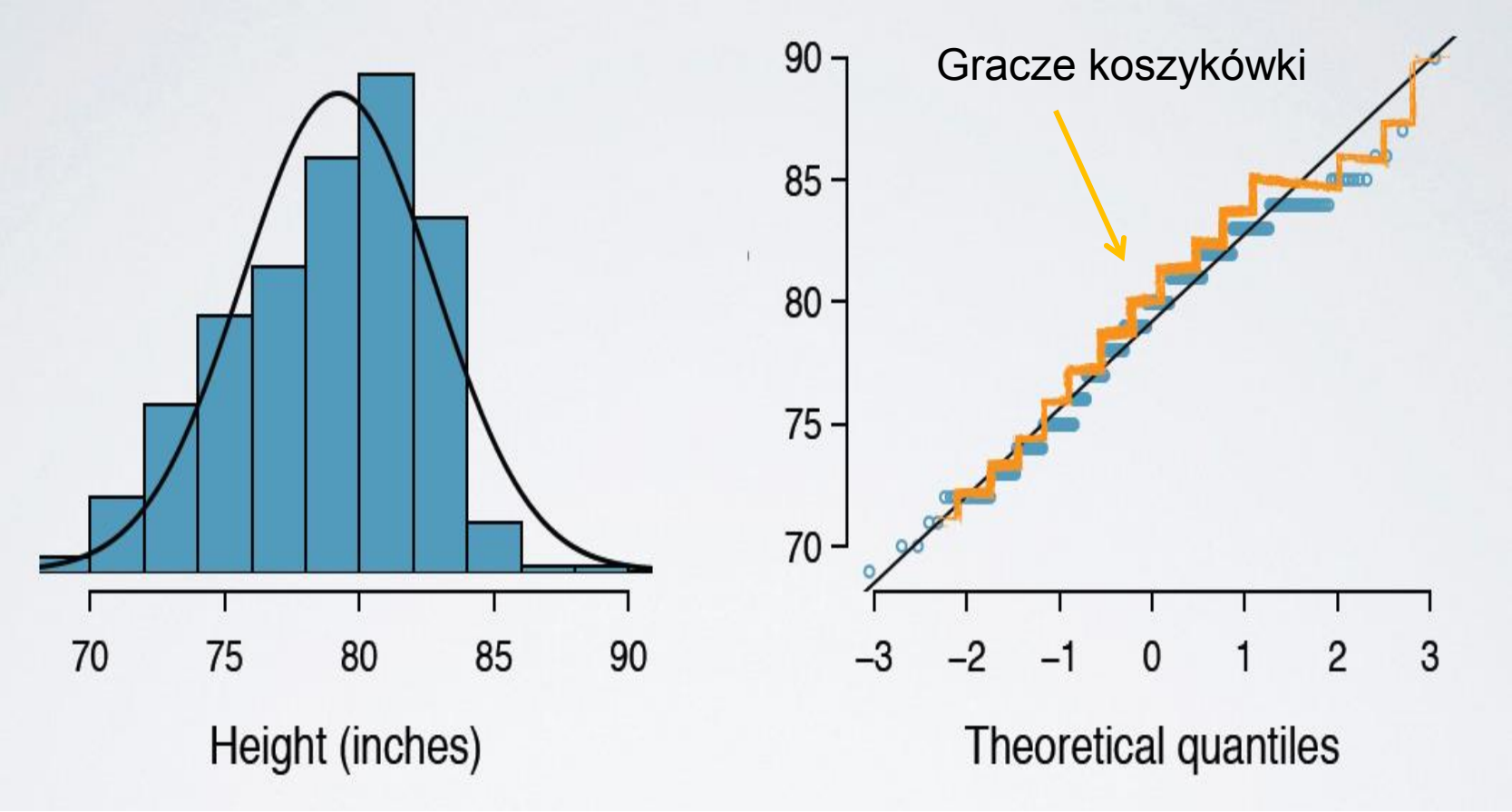
Plot normalnego prawdopodobieństwa



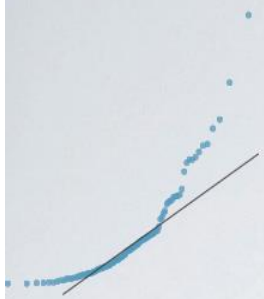
Plot normalnego prawdopodobieństwa

- Dane są na osi pionowej, na osi poziomej są teoretyczne kwantyle (rozkładu normalnego)
- Jeżeli jest dobra zgodność danych i teoretycznych kwantyli to dane układają się na linii prostej
- Im bliżej prostej linii, tym rozkład bliższy rozkładowi normalnemu
- Ten plot wymaga obliczania „percentili” i „z-scores” czyli standaryzowanej zmiennej.

Plot normalnego prawdopodobieństwa



Plot normalnego prawdopodobieństwa



Right skew
Points bend up and to the left of the line.

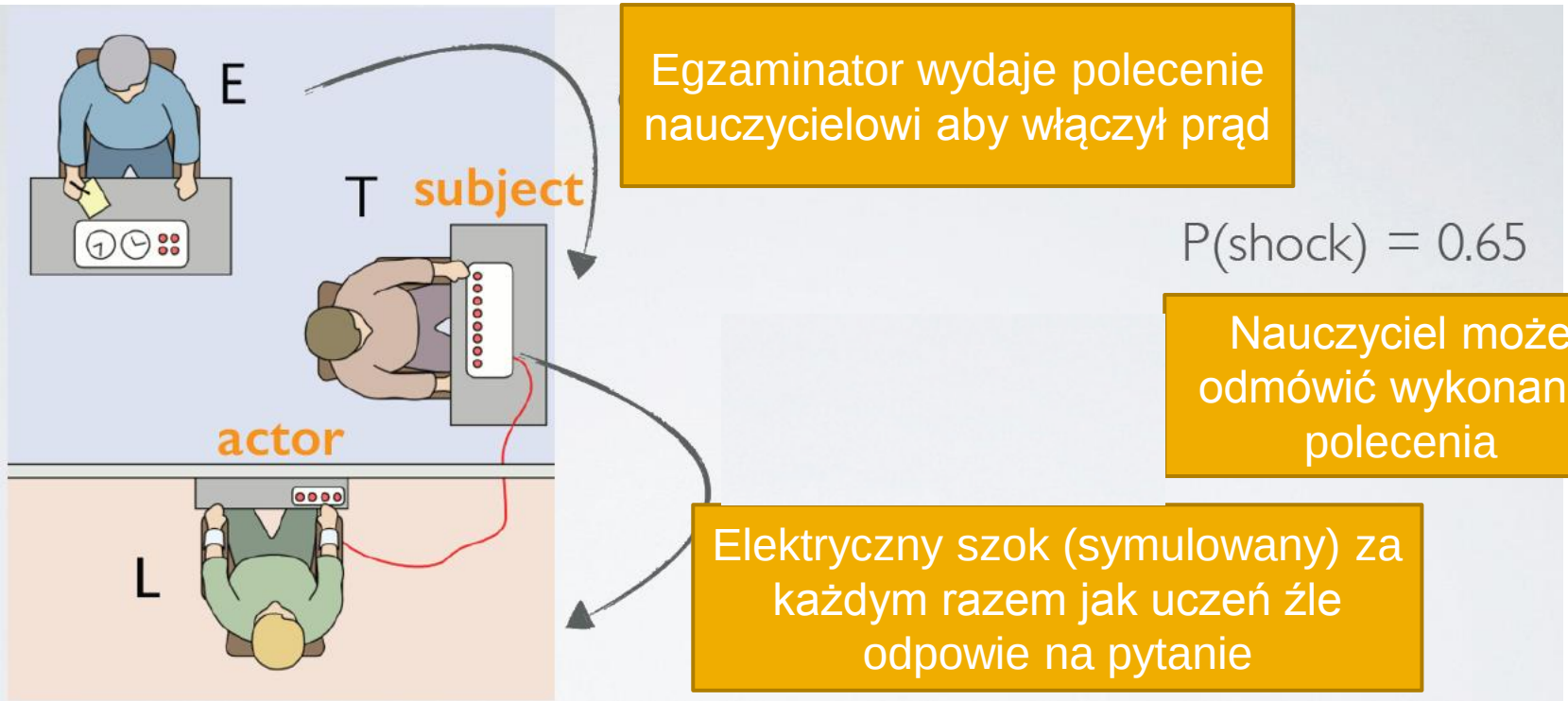
Short tails (narrower than the normal distribution)
Points follow an S shaped-curve.

Left skew
Points bend down and to the right of the line.

Long tails (wider than the normal distribution)
Points start below the line, bend to follow it, and end above it.

Rozkład binomialny

Milgram eksperyment



Rozkład Bernouliego

- Każdy nauczyciel przystępujący do testów jest traktowana jak „próba”
- Jeżeli odmówi wykonania polecenia i włączenia prądu to to jest sukces. Tylko 35% odmawia więc P(sukces) =0.35
- Każda próba ma tylko dwie możliwe wartości: tak lub nie na wydane polecenie. Podlega więc rozkładowi Bernouliego.

Symulacja

Wybieramy losowo 4 osoby aby wzięły udział w eksperymencie. Jakie prawdopodobieństwo że dokładnie 1 odmówi wykonania polecenia.

- Four individuals:
  (A) Anthony
  (B) Brittany
  (C) Clara
  (D) Dorian
- Multiple scenarios where *"exactly 1 refuses"*

Scenario 1: $0.35_{\text{(A) refuse}} \times 0.65_{\text{(B) shock}} \times 0.65_{\text{(C) shock}} \times 0.65_{\text{(D) shock}} = 0.0961$

OR

Scenario 2: $0.65_{\text{(A) shock}} \times 0.35_{\text{(B) refuse}} \times 0.65_{\text{(C) shock}} \times 0.65_{\text{(D) shock}} = 0.0961$

OR

Scenario 3: $0.65_{\text{(A) shock}} \times 0.65_{\text{(B) shock}} \times 0.35_{\text{(C) refuse}} \times 0.65_{\text{(D) shock}} = 0.0961$

OR

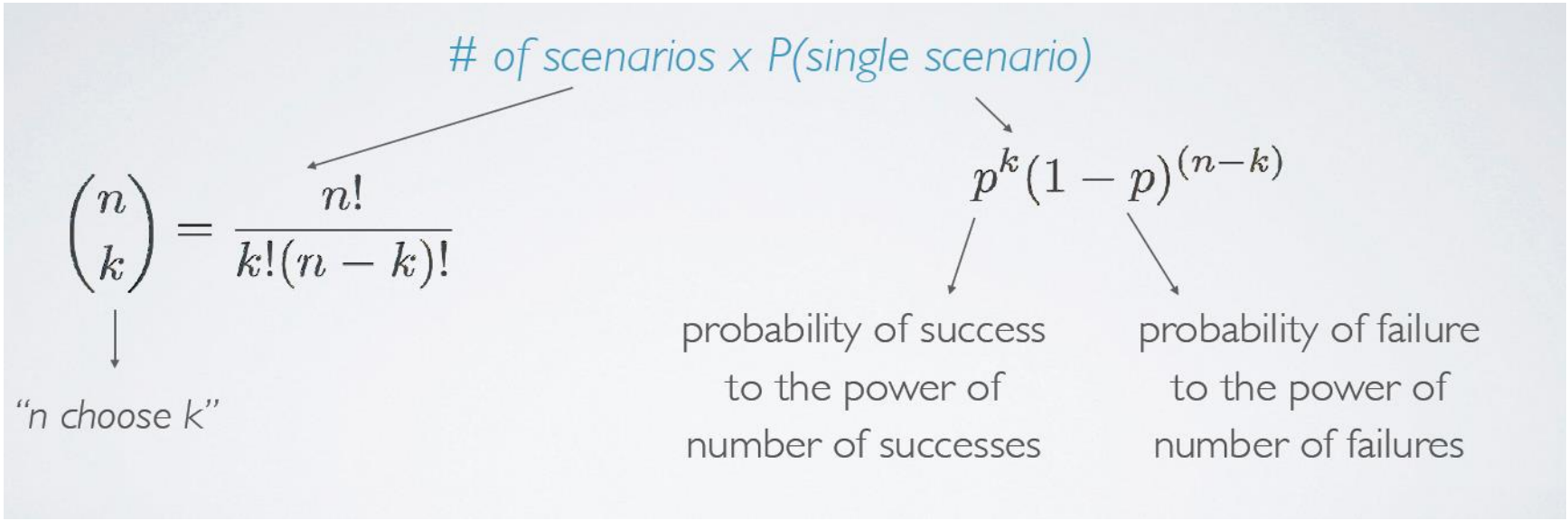
Scenario 4: $0.65_{\text{(A) shock}} \times 0.65_{\text{(B) shock}} \times 0.65_{\text{(C) shock}} \times 0.35_{\text{(D) refuse}} = 0.0961$

$+$

$4 \times 0.0961 = 0.3844$

Rozkład binomialny

Rozkład binomialny to rozkład prawdopodobieństwa dokładnie **k** sukcesów przy n-próbach Bernouliego z prawdopodobieństwem pojedynczego sukcesu **p**



Rozkład binomialny

Ile jest możliwości dla
1 sukcesu w 4 próbach?

$n = 4 \quad k = 1$

$$\binom{4}{1} = \frac{4!}{1! \times (4-1)!} = \frac{4 \times 3 \times 2 \times 1}{1 \times 3 \times 2 \times 1} = 4$$

Ile jest możliwości dla
2 sukcesów w 9 próbach?

SSFFFFFFF
SFSFFFFFF
SFFSFFFFF
...

$n = 9 \quad k = 2$

$$\binom{9}{2} = \frac{9!}{2! \times 7!} = \frac{9 \times 8 \times 7!}{2 \times 1 \times 7!} = 36$$

R

```
> choose(9,2)
[1] 36
```

Rozkład binomialny

Jeżeli p = prawdopodobieństwo sukcesu, (1-p) = prawdopodobieństwo porażki, n = ilość prób, k=ilość sukcesów to:

$$P(\text{k successes in n trials}) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

gdzie

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Warunki które muszą być spełnione:

- Próby są niezależne
- Ilość prób musi być ustalona z góry
- Każda próba musi być sklasyfikowana jako sukces lub porażka
- Prawdopodobieństwo sukcesu musi być to samo dla każdej próby

Przykład

Zgodnie z badaniami statystycznymi (2013), tylko **13% pracujących** jest zaangażowanych w swoja pracę (lubi ją i chce mieć pozytywny wpływ). Wybierając losowo **10 osób**, jakie jest prawdopodobieństwo że **dokładnie 8** jest zaangażowanych w pracę

$n = 10$

$p = 0.13$

$1 - p = 0.87$

$k = 8$

$$P(K = 8) = \binom{10}{8} 0.13^8 \times 0.87^2$$

$$= \frac{10!}{8! \times 2!} \times 0.13^8 \times 0.87^2$$

$$= \frac{10 \times 9 \times 8!}{8! \times 2 \times 1} \times 0.13^8 \times 0.87^2$$

$$= 45 \times 0.13^8 \times 0.87^2$$

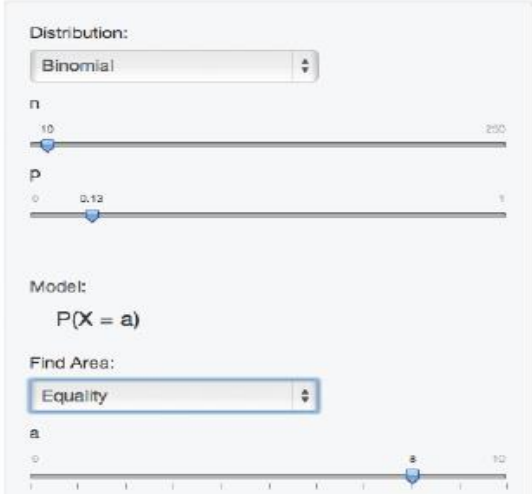
$$= 0.00000278$$

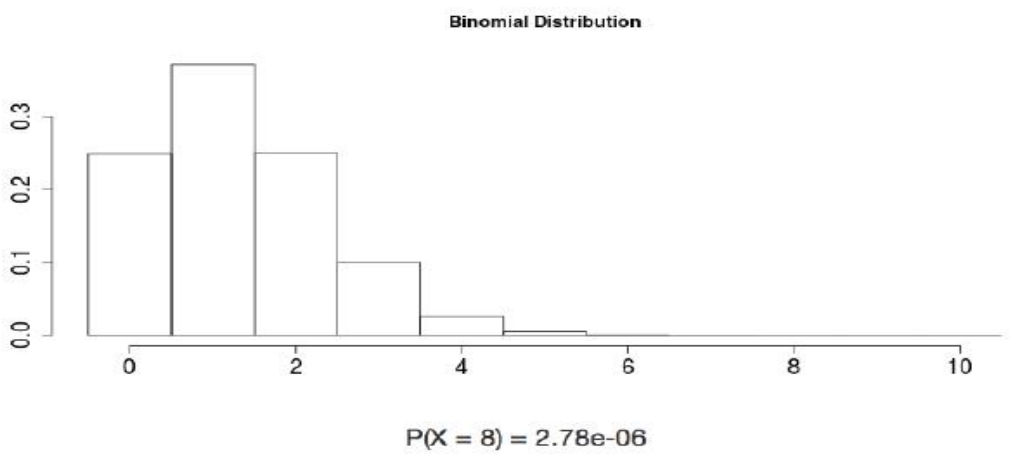
Przykład (cd.)

R

```
> dbinom(8, size = 10, p = 0.13)
[1] 2.77842e-06
```

http://bit.ly/dist_calc





Przykład (cd.)

- Wybierając losowo 100 pracowników, ilu oczekujesz że będzie zaangażowanych w swoja pracę? Pamiętajmy **p=0.13**

$$\mu = np$$

$$\mu = 100 \times 0.13 = 13$$

$$\sigma = \sqrt{np(1-p)}$$

$$\sigma = \sqrt{100 \times 0.13 \times 0.87} = 3.36$$

Przybliżenie rozkładu binomialnego

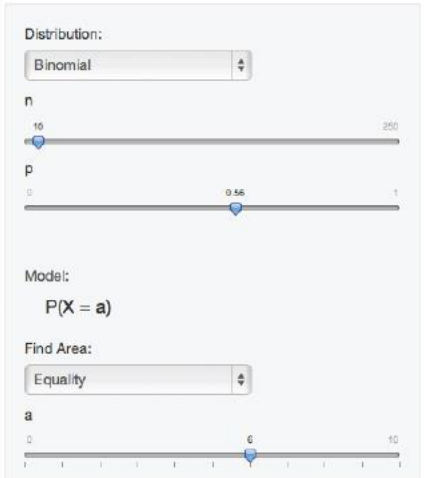
- Rozkład normalny może być dobrym przybliżeniem rozkładu binomialnego.
- Warunek:
  - Próbka jest wystarczająco duża aby

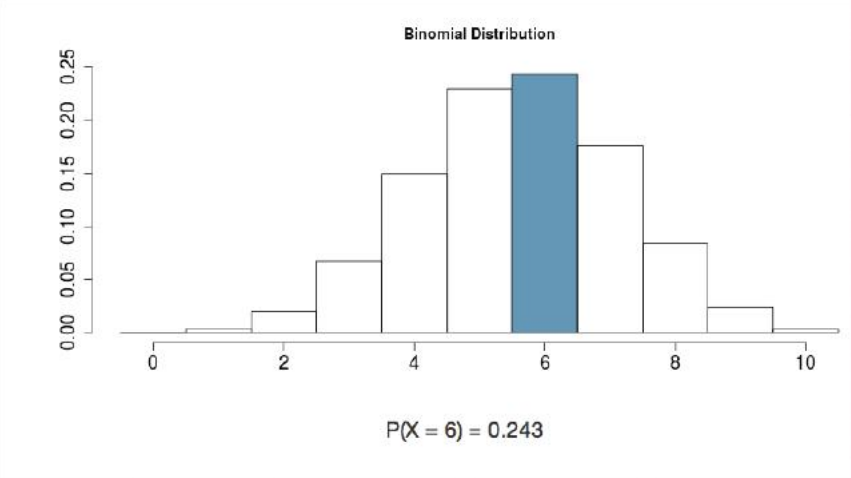
    np > 10 oraz **n(1-p) > 10**

Przykład

Zgodnie z badaniami statystycznymi (2014), 56% obywateli USA którzy nie są ubezpieczeni zamierza się ubezpieczyć korzystając z programu rządowego. Jakie jest prawdopodobieństwo że z 10 losowo wybranych osób, dokładnie 6 zamierza się ubezpieczyć?

http://bit.ly/dist_calc





Przykład (cd)

Zgodnie z badaniami statystycznymi (2014), 56% obywateli USA którzy nie są ubezpieczeni zamierza się ubezpieczyć korzystając z programu rządowego.

Jakie jest prawdopodobieństwo że z 10 losowo wybranych osób, dokładnie 6 zamierza się ubezpieczyć?

R

```
> dbinom(6, size = 10, p = 0.56)
[1] 0.243
```

Przykład (cd)

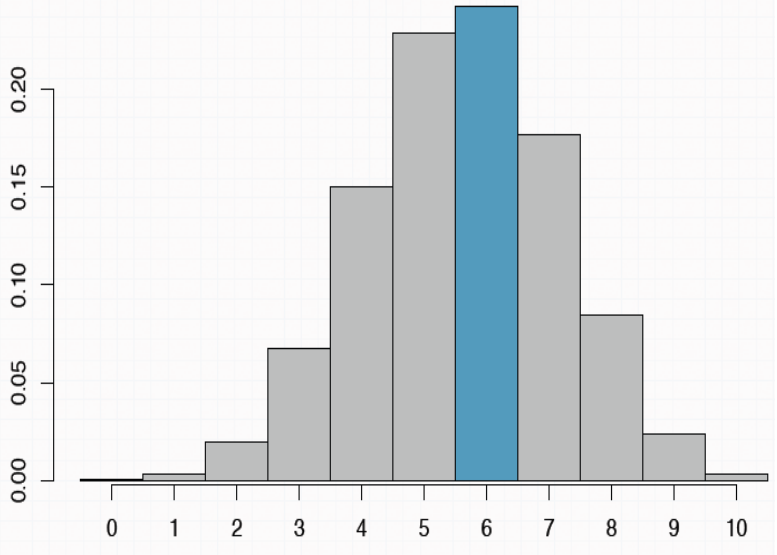
Zgodnie z badaniami statystycznymi (2014), 56% obywateli USA którzy nie są ubezpieczeni zamierza się ubezpieczyć korzystając z programu rządowego. Jakie jest prawdopodobieństwo że z 10 losowo wybranych osób, dokładnie 6 zamierza się ubezpieczyć?

$$P(K=6) = \binom{10}{6} 0.56^6 \times 0.44^4$$

$$= \frac{10 \times 9 \times 8 \times 7 \times 6!}{6! \times 4 \times 3 \times 2 \times 1} \times 0.56^6 \times 0.44^4$$

$$= 210 \times 0.56^6 \times 0.44^4$$

$$= 0.243$$



Przykład (cd)

Jakie jest prawdopodobieństwo że
z 1000 losowo wybranych osób, dokładnie 600
zamierza się ubezpieczyć?

(a) 0.243, same as P(K = 6)
(b) less than 0.243
(c) more than 0.243

$p = 0.56$

$n_1 = 10 \qquad \mu_1 = 10 \times 0.56 = 5.6$

$\Delta = 6 - 5.6 = 0.4$

$n_2 = 1000 \quad \mu_2 = 1000 \times 0.56 = 560$

$\Delta = 600 - 560 = 40$

R

```
> dbinom(600, 1000, 0.56)

[1] 0.00098
```

Przykład (cd)

Jaki jest rozkład prawdopodobieństwa osób które planują się ubezpieczyć w pośród 100 losowo wybranych osób

$p = 0.56 \quad n = 100$

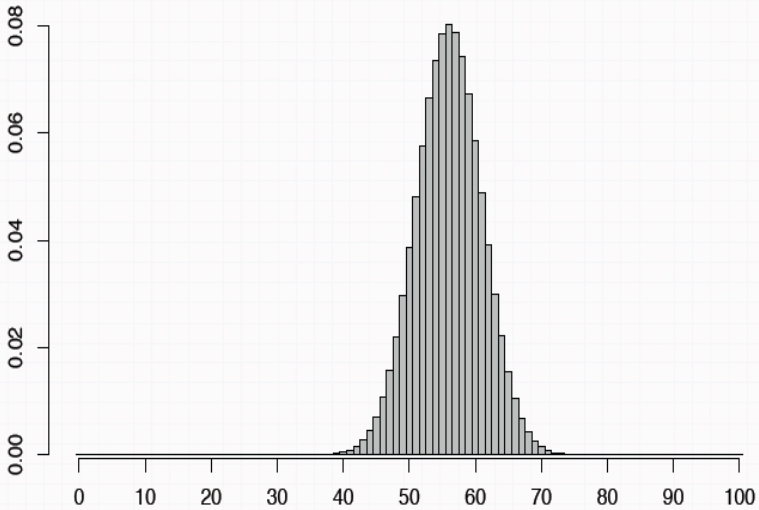
$\#\ suc: np = 100 \times 0.56 = 56 > 10$

$\#\ fail: n(1-p) = 100 \times 0.44 = 44 > 10$

$\mu = 56$

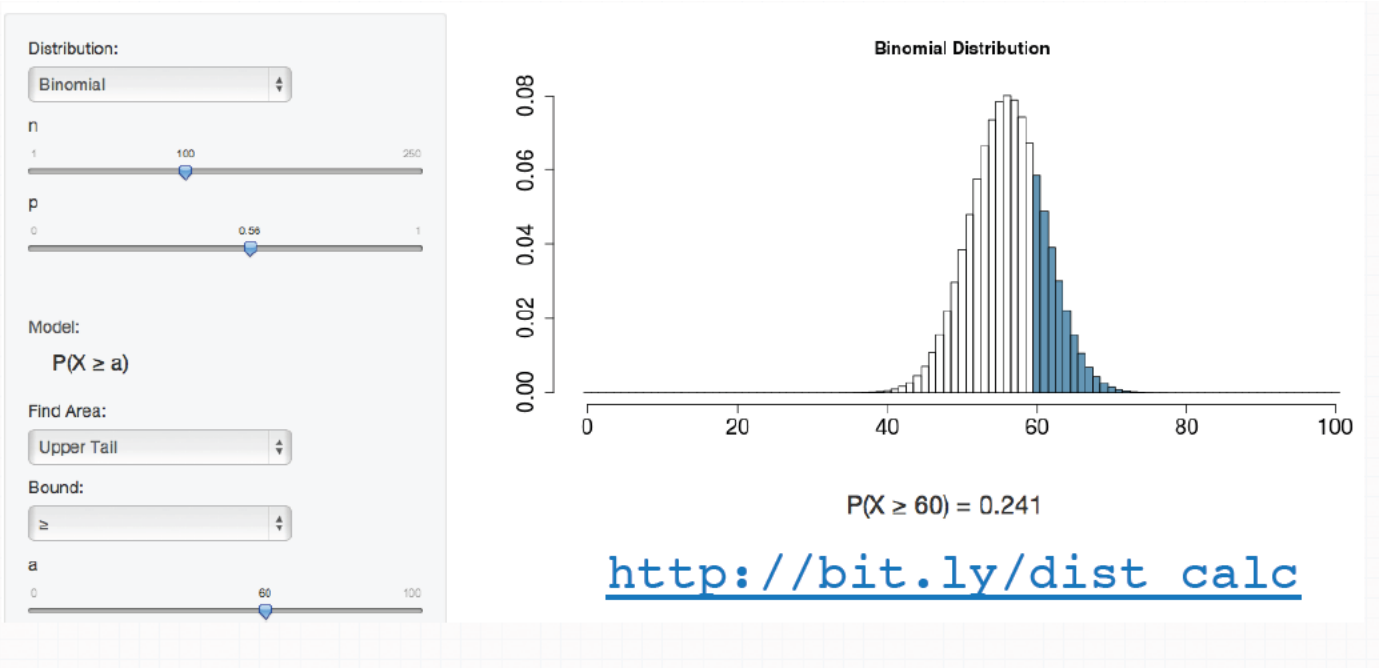
$\sigma = \sqrt{100 \times 0.56 \times 0.44} = 4.96$

$N(mean = 56, SD = 4.96)$



Przykład (cd)

Jaki jest rozkład prawdopodobieństwo że co najmniej 60 osób spośród 100 losowo wybranych planuje się ubezpieczyć.



Przykład (cd)

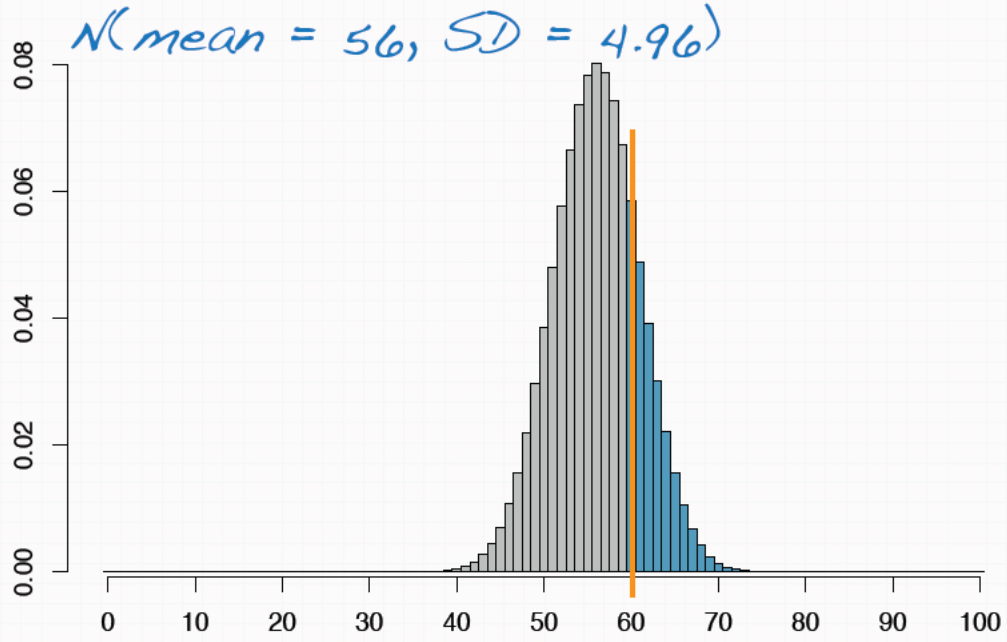
Jaki jest rozkład prawdopodobieństwo że co najmniej 60 osób spośród 100 losowo wybranych planuje się ubezpieczyć.

R

```
> sum(dbinom(60:100, size = 100, p = 0.56))
[1] 0.241
```

Przykład (cd)

Jaki jest rozkład prawdopodobieństwo że co najmniej 60 osób spośród 100 losowo wybranych planuje się ubezpieczyć.



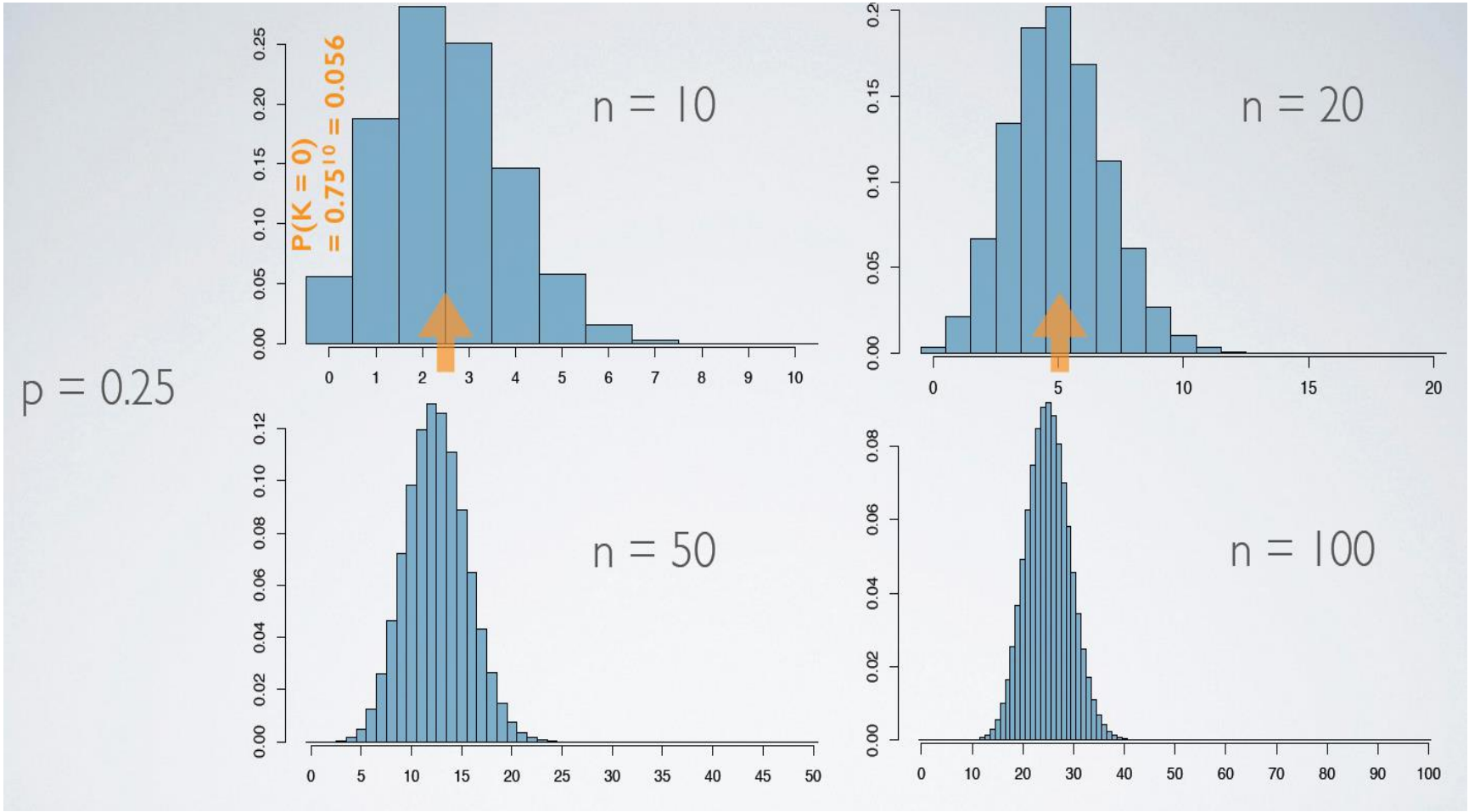
$$Z = \frac{60 - 56}{4.96} \approx 0.81$$

$$P(Z > 0.81) = 0.209$$

$$Z = \frac{59.5 - 56}{4.96} \approx 0.71$$

$$P(Z > 0.71) = 0.239$$

Rzkład binomialny → normalny



Rzkład binomialny → normalny

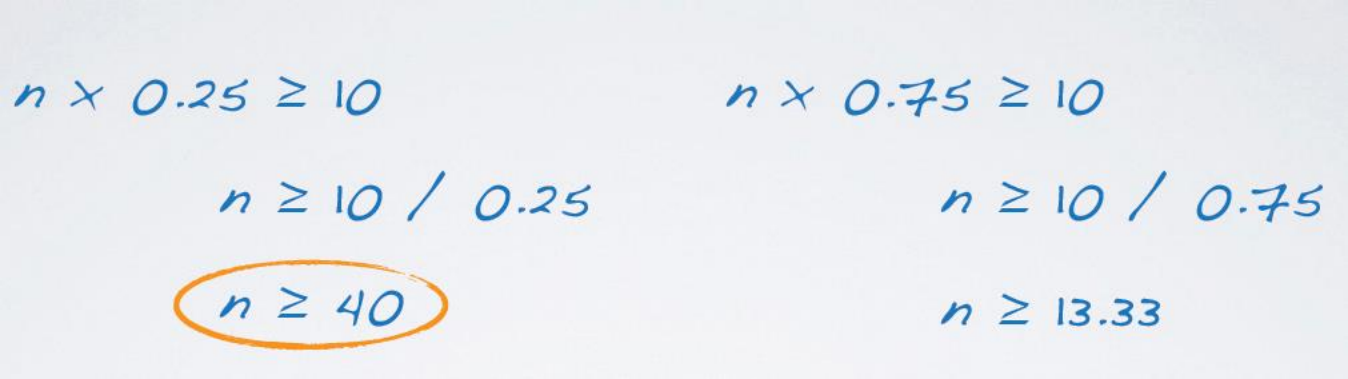
- **Sukces-przegrana reguła:**
  - rozkład binomialny który ma co najmniej 10 oczekiwanych sukcesów i 10 oczekiwanych porażek jest blisko rozkładu normalnego.
- Jeżeli to co powyżej jest spełnione to

$$\text{Binomial}(n,p) \sim \text{Normal}(\mu,\sigma)$$

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

Rzkład binomialny → normalny

- Jaka jest minimalna próbka **n** o rozkładzie binomialnym aby była blisko rozkładu normalnego?

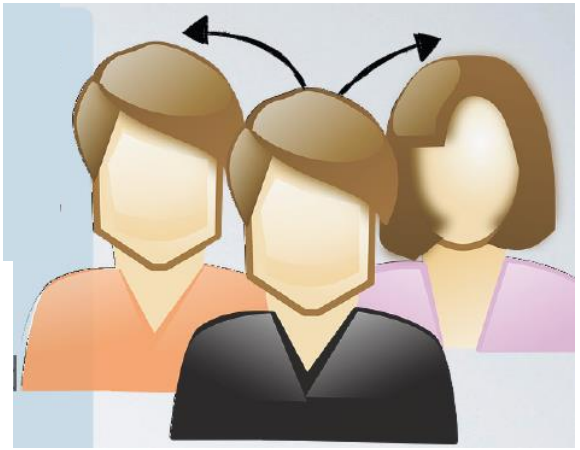


Przykład

Średni użytkownik facebook „dostaje więcej niż daje”

- „szukam przyjaciela”: 40% wysłali i 63% dostali
- „lubię” : 14 x wysłali, 20x dostali (średnio)
- „wiadomość”: 9 wysłali i 12 dostali (średnio)
- „tags”: 12% znalazło a 35% było znalezionych

- Inne:
  - 25% uważanych jest za „power user”
  - Średni użytkownik facebook ma 245 przyjaciół
- Jakie jest prawdopodobieństwo że średni użytkownik ma co najmiej 70 przyjaciół którzy są „power user”?



$p = 0.25$

$n = 245$

$P(K \geq 70) = ?$

Image: http://pixabay.com/p-35681/?no_redirect

Poll: http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

Przykład (cd)

$P(K \geq 70) = ?$

$= P(K = 70) + P(K = 71) + \ldots + P(K = 245)$

(1) $n = 245$, fixed

(2) power user / not

(3) $p = 0.25$

(4) independence

$N(\text{mean}, SD)$

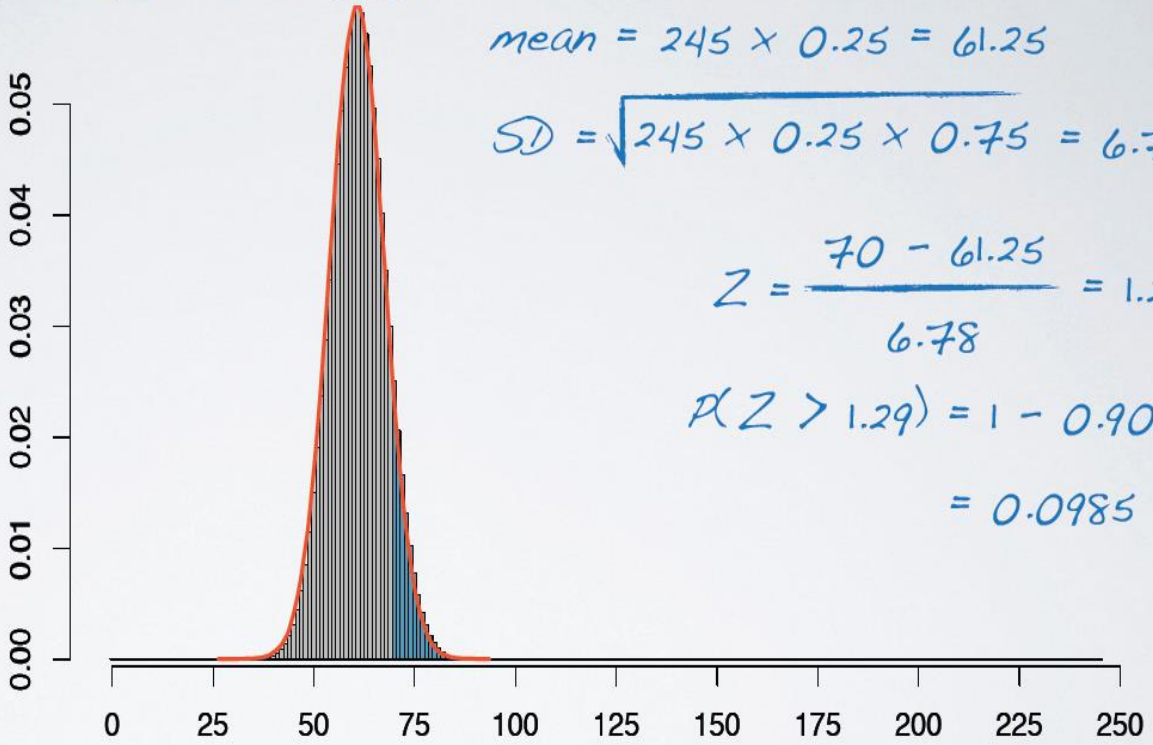
$\text{mean} = 245 \times 0.25 = 61.25$

$SD = \sqrt{245 \times 0.25 \times 0.75} = 6.78$

$$Z = \frac{70 - 61.25}{6.78} = 1.29$$

$P(Z > 1.29) = 1 - 0.9015$

$= 0.0985$



Przykład (cd)

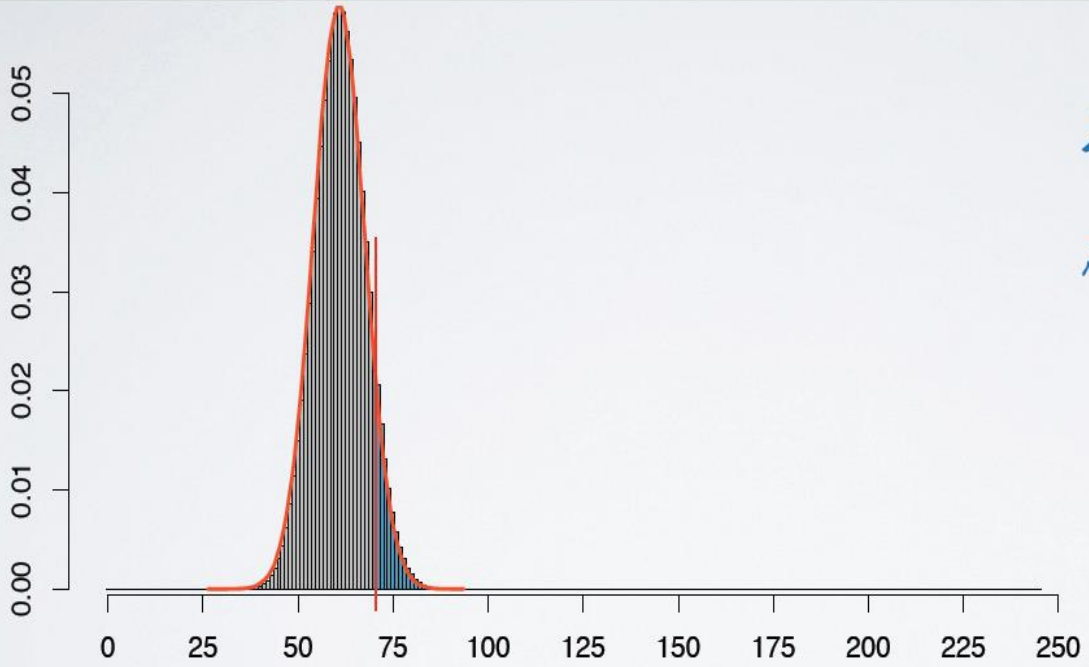
	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

Przykład (cd)

R

```
> sum(dbinom(70:245, size = 245, p = 0.25))
[1] 0.113
```

vs 0.0985???



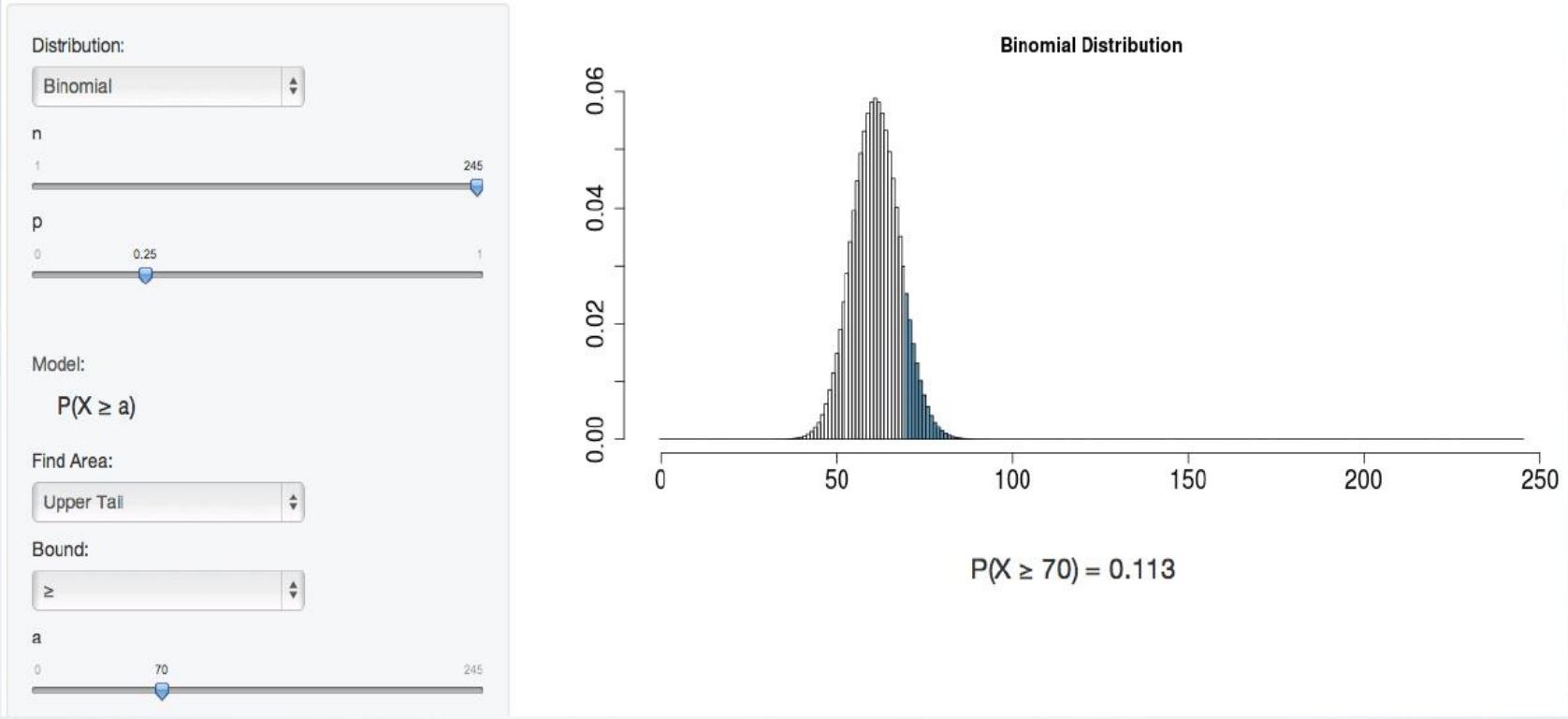
$$Z = \frac{69.5 - 61.25}{6.78} = 1.22$$

$$P(Z > 1.22) = 1 - 0.8888$$

$$= 0.1112$$

Przykład (cd)

http://bit.ly/dist_calc



Negatywny rozkład binomialny

- Obserwacja dokładnie k-tego sukcesu w n-tej próbie.
- Warunki:
  - Każda próba jest niezależna
  - Wynik każdej próby jest sukcesem albo porażką
  - Prawdopodobieństwo sukcesu jest takie same dla każdej próby
  - Ostatnia próba kończy się sukcesem

Negatywny rozkład binomialny

Negative binomial distribution

The negative binomial distribution describes the probability of observing the k^{th} success on the n^{th} trial:

$$P(\text{the } k^{th} \text{ success on the } n^{th} \text{ trial}) = \binom{n-1}{k-1} p^k (1-p)^{n-k} \qquad (3.58)$$

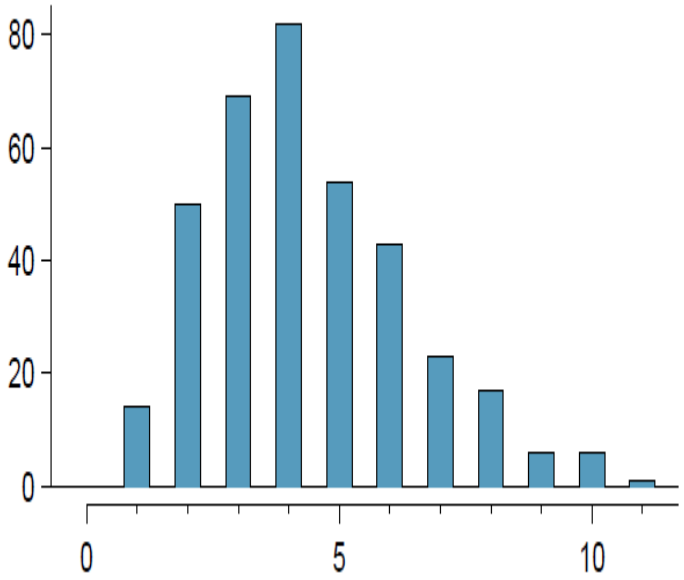
where p is the probability an individual trial is a success. All trials are assumed to be independent.

Binomialny vs Negatywny binomialny

- Binomialny:
  - Ustalona ilość prób
  - Rozważamy ilość sukcesów
- Negatywny binomialny
  - Pytamy jak wiele prób należy wykonać aby zaobserwować k sukcesów, żądamy aby ostatnia próba kończyła się sukcesem

Rozkład Poissona

- Zmienna losowa ma taki rozkład jeżeli przypadek zachodzi rzadko, cała próbka jest duża i poszczególne zdarzenia mogą być uważane za niezależne.
- Rzadkie zdarzenia np.:
  - Atak serca
  - Wzięcie ślubu
  - Uderzenie pioruna



Rozkład Poissona

Poisson distribution

Suppose we are watching for rare events and the number of observed events follows a Poisson distribution with rate λ. Then

$$P(\text{observe } k \text{ rare events}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where k may take a value 0, 1, 2, and so on, and $k!$ represents k-factorial, as described on page 138. The letter $e \approx 2.718$ is the base of the natural logarithm. The mean and standard deviation of this distribution are λ and $\sqrt{\lambda}$, respectively.