

Wykład 2

- Wpływ przekształceń
- Co się stanie ze średnią i odchyleniem standardowym gdy zmienimy jednostki ?
 - Stopnie Celsiusza $\leftrightarrow$ stopnie Fahrenheita
 - dolary $\leftrightarrow$ 1,000 dolarów
 - Wartość faktyczna $\leftrightarrow$ odległość od minimum
 - cm to mm, in, nm, m, ft, itd. dolary : euro

- Zmiana jednostek na ogół da się zapisać za pomocą funkcji liniowej

$$y' = ay + c$$

Np.

- $y' = 1.8 y + 32$
- $y' = 1/1000 y (+ 0)$
- $y' = (1)y - y_{\min}$

- Czasami $a = 1$ or $c = 0$
- Współczynnik kierunkowy a
- Stała c
- Funkcja liniowa nie zmienia w zasadniczy sposób kształtu histogramu. Może go rozszerzyć ($|a|$), „ściśnić” ($|a|$), przesunąć (c) i „obrócić” (a).

- Średnia
- $\bar{y}$ zmienia się tak jak y tzn.
- $\bar{y}' = a \bar{y} + c$
- Odchylenie standardowe
- Odchylenie standardowe zależy od współczynnika kierunkowego a . Stała c nie ma wpływu na odchylenie standardowe, ponieważ zależy ono jedynie od odchylenia od średniej.
- $s' = |a| s$

Wpływ stałej (odejmujemy 20)

	y	Dev.	y'	Dev
	25	-1	5	-1
	26	0	6	0
	28	2	8	2
	25	-1	5	-1
Mean	26		6	

- Wariancja
- Wariancję mnożymy przez a^2 .
 $s'^2 = a^2 s^2$
- Przykład
- Y- temperatura w °F: $\bar{y} = 98.6$, $s = 0.9$, $s^2 = 0.81$
- Pytanie 1: Oblicz średnią, odchylenie standardowe i wariancję dla tych samych danych wyrażonych w stopniach Celsiusza.

Odpowiedź

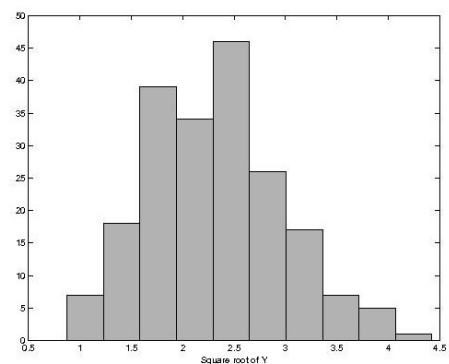
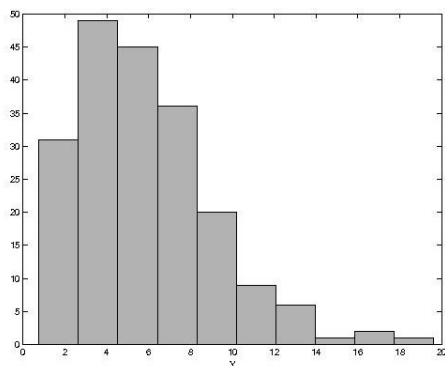
Standardyzacja

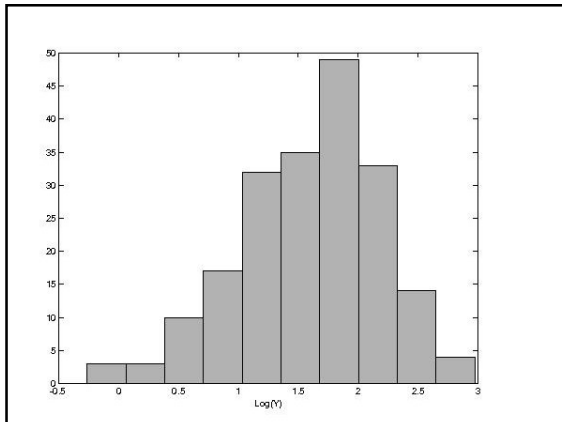
- Pytanie 2: Jakich wyników należy oczekiwać gdy dane przekształcimy w następujący sposób
- $Y' = (Y - \bar{Y})/s$,
tzn. $Y' = 1/s Y - \bar{y}/s = (Y - 98.6)/1.8$?
- Odpowiedź:

Inne statystyki

- Funkcja liniowa zmienia medianę i kwartyle tak jak średnią, a rozstęp i IQR tak jak odchylenie standardowe.

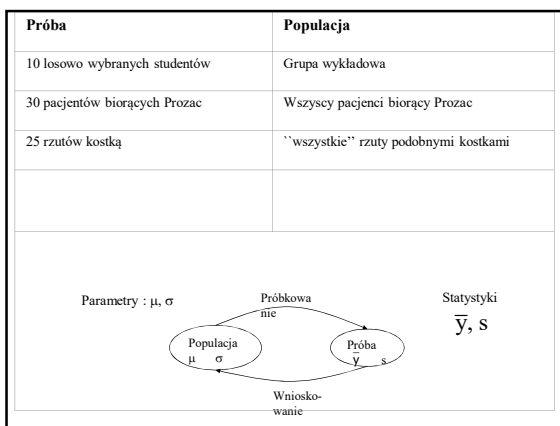
- Funkcje nieliniowe (np. logarytm) zmieniają kształt histogramu i na ogół nie ma dla nich prostych formuł umożliwiających obliczenie nowej średniej i nowego odchylenia standardowego. Parametry te liczymy z definicji korzystając z "nowego" zbioru danych.
- Przykład : $Y' = \log(Y)$
średnia z $y' \neq \log(\text{średnia z } y)$
- Czasami używamy funkcji nieliniowych aby przekształcić skośny zbiór danych w zbiór bardziej symetryczny.





Próba a populacja : Wnioskowanie statystyczne

- Co to jest populacja ?
 - Zbiór z którego losujemy próbę i który chcemy opisać.
 - Czasami rzeczywista, czasami abstrakcyjna (nieskończenie duża próba) .
 - Próba powinna być reprezentatywna dla populacji.
- Wnioskowanie statystyczne
 - Wnioskowanie o populacji w oparciu o próbę.



- μ = średnia w populacji (wartość oczekiwana)
- σ = odchylenie standardowe w populacji = pierwiastek z wartości oczekiwanej dla $(Y-\mu)^2$
- Statystyki z próby są *estymatorami* (służą do oceny) parametrów całej populacji.

Przykład

- Grupy krwi u 3696 osób (Anglia)

Grupa krwi	Częstość
A	1,634
B	327
AB	119
O	1616
suma	3696

- W przybliżeniu 44% ludzi ma grupę krwi A.
 - Czy nie było systematycznego błędu przy próbkowaniu?
 - Czy rozmiar próby był dość duży?

- Próba : przyjaciele, pracownicy
- Ludzie nie są reprezentatywni
- Grupy krwi mogą być mimo to reprezentatywne
- Pomiary ciśnienia nie byłyby reprezentatywne (w próbie mało starszych ludzi – ciśnienie na ogół wzrasta z wiekiem)

Populacja a próba

- Średnia z próby na ogół różni się od wartości oczekiwanej (średniej w populacji) ale w miarę wzrostu rozmiaru próby odległość między tymi wielkościami zbiega do zera.
- Średnia z próby jest estymatorem wartości oczekiwanej. Podobnie próbkowe odchylenie standardowe i wariancja są estymatorami odpowiednich parametrów w populacji.
 - Przykład (rozmiar populacji=50, średnia w populacji =26.48)
 dane : 25.5 17.8 36.7 29.8 40.7 26.0 7.7 27.7 10.3 22.3 45.4
 43.4 20.2 42.2 44.5 1.6 5.7 48.6 23.9 27.2 17.0 19.5 47.7 3.9
 39.3 9.2 30.7 18.9 25.7 32.8 16.8 11.7 13.9 4.9 49.4 30.5 20.7
 38.1 25.6 40.7 45.0 30.8 11.3 34.0 49.7 21.3 3.5 28.7 19.7 35.6
 - stopniowo powiększamy próbę losową do rozmiarów $n=10, 20, 30, 40$
 średnia z próby : 23.5 ($n=10$), 27.3 ($n=20$), 26.7 ($n=30$), 26.4 ($n=40$)

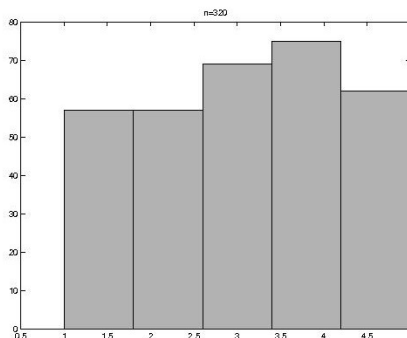
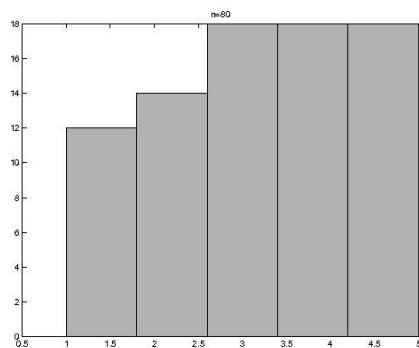
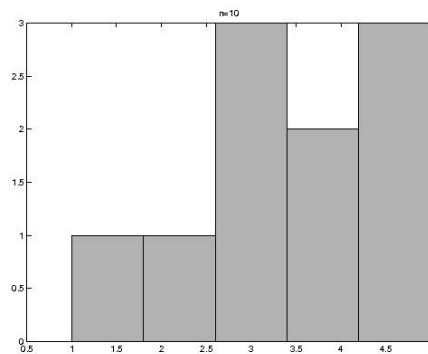
Histogram z populacji i histogram próbkowy

- Dane dyskretne (klasy)
 - ✓ p_i -frakcja osobników w całej populacji kwalifikujących się do kategorii i
 - ✓ Można ustalić p_i w oparciu o histogram skonstruowany dla całej populacji.
 - ✓ $\hat{p}_i$ - estymator obliczony w oparciu o histogram z próby.

Przykład

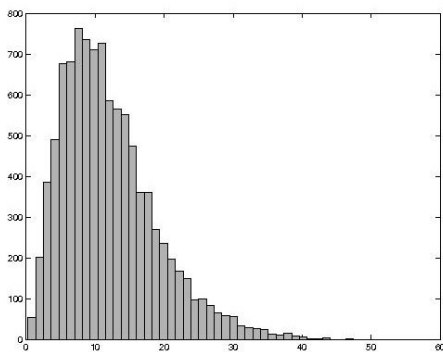
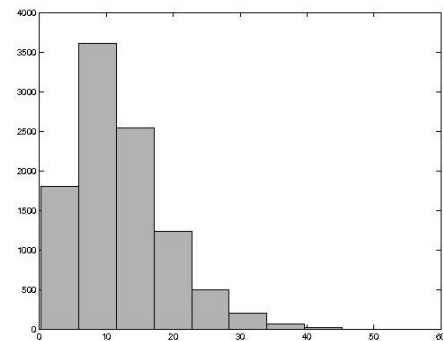
Rozmiar populacji =10000. 5 klas o tej samej częstości $p=?$.

n	$\hat{p}_1$	$\hat{p}_5$
10	0.1	0.3
20	0.1	0.35
40	0.2	0.25
80	0.15	0.225
160	0.1625	0.1875
320	0.1781	0.1938



Histogramy a gęstości rozkładu prawdopodobieństwa

- Dane ciągłe
- Więcej danych – więcej klas – gładzy histogram



Prawdopodobieństwo

- Gdy rozmiar próby się zwiększa a szerokość klas się zmniejsza to histogram zbiega do wykresu gęstości rozkładu badanej zmiennej w populacji.
- Podobnie jak dla histogramu, pole pod wykresem gęstości (całka) jest równe frakcji osobników w całej populacji wpadających do danego przedziału.
- To pole jest też równe p-stwu, że losowo wybrany osobnik będzie w zadanym przedziale.

Gęstość

- Gęstość rozkładu prawdopodobieństwa spełnia następujące warunki:
 - $f(x) \geq 0$ dla wszystkich x
 - Całkowite pole pod wykresem $f(x)$ wynosi ?

$$\int_{-\infty}^{\infty} f(x) dx = ?$$

Przykłady rozkładów ciągłych

- Rozkład jednostajny na odcinku $[a, b]$
- $f(x) = ?$
- Rozkład wykładniczy z parametrem λ
- $f(x) = ?$

Rozkłady ciągłe cd.

- Rozkłady ciągłe określają prawdopodobieństwa tego, że obserwacje wpadają do pewnych odcinków

$$P(X \in (a, b)) = \int_a^b f(x) dx$$

- Prawdopodobieństwo przyjęcia konkretnej dowolnej wartości przez rozkład ciągły wynosi ?
- Niech Y ma rozkład jednostajny na odcinku [0,1].
- $P(Y=0.5)=?$

Próbkowanie

- Prosta próba losowa
 - Każdy osobnik z populacji może być wybrany z tym samym prawdopodobieństwem
 - Wybory poszczególnych osobników są od siebie niezależne

- Jak wybrać prostą próbę losową
 - Przyznajemy numer każdemu osobnikowi
 - Zapisujemy numery na kulach
 - Mieszmamy kule w urnie
 - Losujemy próbę
- Do losowania możemy również użyć komputera lub tablicę liczb losowych.
 - W przypadku gdy rozmiar populacji nie jest ustalony lub nie mamy bezpośredniego dostępu do wszystkich osobników zadanie jest dużo trudniejsze.

Przykład 1(Ochotnicy)

- Dziennikarka Ann Landers spytała swoich czytelników ``Gdybyście mogli zacząć jeszcze raz to czy mielibyście znowu dzieci ? ``
- Odpisało prawie 10,000 czytelników i 70% powiedziało Nie!
- Obserwacje: czytelnicy pisma Anny Landers
- Populacja: wszyscy rodzice w USA

- Ochotnicy: badanie bezwartościowe – bardzo zła reprezentatywność
- *Newsday przeprowadził ``statystycznie zaplanowaną`` ankietę, w której 91% z 1,373 przepytanych rodziców odpowiedziało Tak!*

Przykład 2 (U.S. przewidywanie wyników wyborów prezydenckich w USA, 1936)

- Literary Digest wysłał kwestionariusze do 10 milionów ludzi (25% głosujących)
- Odpowiedziało 2.4 miliona:
 - Przewidywanie : Landon 57%, Roosevelt 43%
 - Wynik : Roosevelt 62%, Landon 38%

- Złe próbkowanie
 - Książki telefoniczne, członkostwo klubów, listy zamówień pocztowych, listy właścicieli pojazdów
- Brak odpowiedzi
 - Tylko 24% odpowiedziało (więcej Republikanów)

Obciążenie w próbkowaniu

- Obciążenie w próbkowaniu występuje gdy mamy do czynienia z systematycznym błędem faworyzującym pewną część populacji. W przypadku takiego obciążenia nie pomoże duży rozmiar próby.
- Losowy wybór elementów do próby zwykle eliminuje takie obciążenie.

Stratyfikacja

- Dzielimy populację na podpopulacje podobnych jednostek (warstwy) i oddzielnie próbujemy w każdej warstwie.
- Przykład: studenci & studentki

Próbkowanie wielostopniowe

- Przykład: Badanie w USA dotyczące struktury zatrudnienia. Ankietuje się około 60.000 gospodarstw domowych co miesiąc.
 - Poziom 1: losowa próba z 3,000 counties (województw?)
 - Poziom 2: losowa próba reprezentująca powiaty w każdym wybranym "county"
 - Poziom 3: losowa próba reprezentująca gminy w każdym wybranym powiecie
 - Poziom 3: losowa próba gospodarstw domowych w każdej wybranej gminie

Prawdopodobieństwo- formalizm matematyczny

- A, B, E – zdarzenia losowe
- $P(E)$ – prawdopodobieństwo zdarzenia E
- $0 \leq P(E) \leq 1$
- S – przestrzeń probabilistyczna (zbiór wszystkich możliwych wyników eksperymentu, zawiera wszystkie zdarzenia losowe)
- $P(S) = 1$

$A \cup B$ -

$A \cap B$ -

Jeżeli $A \cap B = \emptyset$

to

$P(A \cup B) =$

- W praktyce prawdopodobieństwa często ustalamy obserwując jaka część populacji posiada interesującą nas własność.
- Przykład – Z 45 studentów 15 dostało 5 z egzaminu. Jakie jest prawdopodobieństwo, że losując studenta z tej grupy trafimy na takiego, który dostał 5 z egzaminu ?

Prawdopodobieństwo klasyczne

- Założenie – wszystkie możliwe wyniki eksperymentu są jednakowo prawdopodobne (prawdopodobieństwo wylosowania każdego studenta jest takie samo).
- N – liczba możliwych wyników eksperymentu ($N = ?$)
- x – liczba tych wyników które odpowiadają zdarzeniu E ($E = 5$ z egzaminu, $x = ?$)
- $P(E) =$

Mocne prawo wielkich liczb

- Gdy liczba powtórzeń eksperymentu dąży do nieskończoności to względna częstość występowania zdarzenia E dąży do $P(E)$.

Przykłady zdarzeń

- E = wyrzucenie orła w rzucie symetryczną monetą
 $P(E) =$
- E = wyrzucenie 4 w rzucie symetryczną kostką: $P(E) =$
- E = 1 lub 6 w rzucie kostką: $P(E) =$

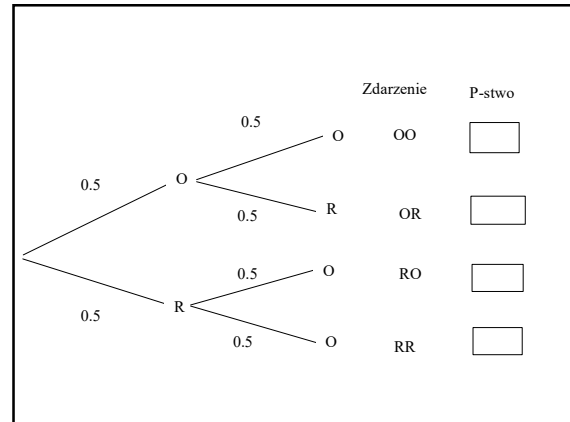
- Ania i Basia rzucają monetą. E = obie dostaną orła.
 $P(E) =$
- Uzasadnienie
 - A dostanie 0 i B dostanie 0
 - A dostanie 0 i B dostanie R
 - A dostanie R i B dostanie 0
 - A dostanie R i B dostanie R
 - Te cztery zdarzenia są jednakowo prawdopodobne ($P(E) = P(00) = P(0R) = P(R0) = P(RR) =$)
 - Prawdopodobieństwo, że dostaniemy dokładnie jednego orła (Ania albo Basia) =

- Krzyżówka dwóch heterozygot
- Genotyp obu rodziców - Aa
- Dzieci: $P(AA) =$, $Pr(Aa \text{ albo } aA) =$, $P(aa) =$
- Jeżeli liczba dzieci będzie bardzo duża to frakcja heterozygot będzie bliska (frakcja w próbie a frakcja w populacji)

Niezależność

Zdarzenia E1 i E2 są niezależne wtedy i tylko wtedy gdy $P(E1 \cap E2) = P(E1) \cdot P(E2)$

- Przykład : dwa rzuty monetą
- $P(\text{dwa orły}) =$



P-stwo warunkowe

$P(A|B)$ – p-stwo zajścia zdarzenia A jeżeli wiemy, że zaszło zdarzenie B

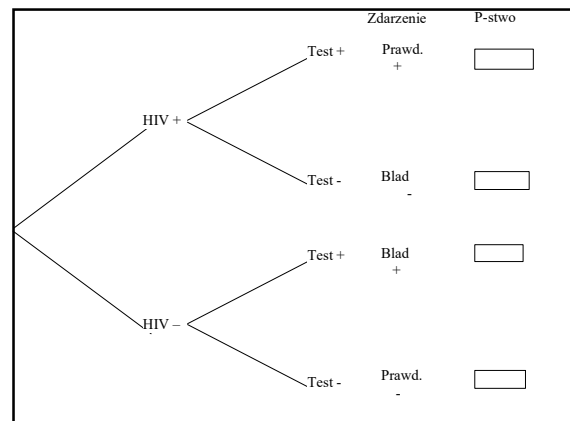
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(B) \cdot P(A|B)$$

Przykład

- 2% populacji ma wirusa HIV.
- Test do wykrywania obecności wirusa HIV ma następujące własności. Jeżeli ktoś ma wirusa to p-stwo jego wykrycia wynosi 0.997. Jeżeli ktoś nie ma wirusa to p-stwo właściwej diagnozy wynosi 0.985.

- A – wybrany losowo człowiek jest chory
- B – test wykazuje obecność wirusa
- $P(A) =$
- $P(B|A) =$
- A' – wybrany losowo człowiek jest zdrowy
- B' – test nie wykazuje obecności wirusa
- $P(B'|A') =$



- Jakie jest p-stwo, że u losowo wybranej osoby test wykaże obecność wirusa ?

- Jakie jest p-stwo, że osoba u której test wskazał obecność wirusa jest faktycznie zakażona ?

Wzór Bayesa

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$