Wykład 4

Plan: 1. Aproksymacja rozkładu dwumianowego rozkładem normalnym

2. Rozkłady próbkowe

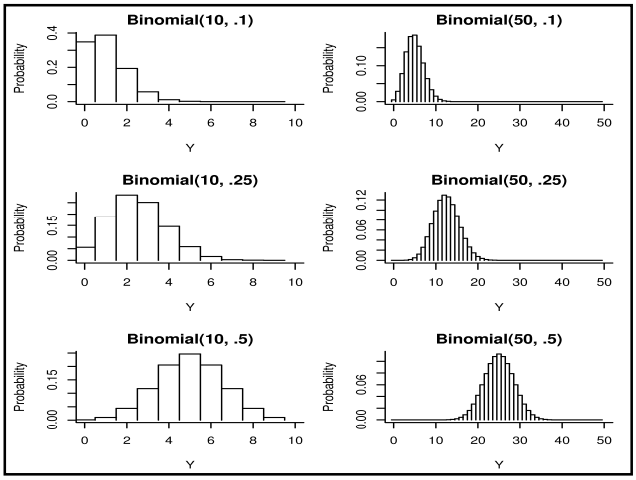
3. Centralne twierdzenie graniczne

Przybliżenie rozkładu *dwumianowego* rozkładem *normalnym*

Niech Y ma rozkład B(n,p).

Liczenie prawdopodobieństw jest problematyczne dla dużych n:

- Gdy **n** jest duże, to mamy dużo wartości **Y**.
- Gdy **n** jest duże, to symbole Newtona są trudne do wyliczenia.



Aproksymacja:

- Rozkład dwumianowy z parametrami *n* i *p* możemy przybliżać rozkładem normalnym z parametrami $\mu=np$ i $\sigma=\sqrt{np(1-p)}$.
- Krótko: *liczymy* jak dla rozkładu normalnego.

Uwagi:

- Przybliżenie dobre, gdy $np \geq 5$ i $n(1-p) \geq 5$.
- Przybliżenie dość dokładne w centrum rozkładu i względnie gorsze w „ogonach”.

Przykład

- Załóżmy, że Y ma rozkład dwumianowy z n=40 i p=0.25.
  Wtedy $\mu = np$ = ... i σ = ...
  Sprawdzamy wielkość np oraz $n(1-p)$:
- Przybliżymy p-stwo tego, że Y jest pomiędzy 10 i 15.

z_1= z_2=

P(10≤Y≤15)≈

[Wynik dokładny z rozkładu dwumianowego = 0.534.]

Korekta na ciągłość

Dość często używamy rozkładu normalnego do opisu danych, które nie są ciągłe. Dokładnosc przybliżenia możemy zwykle poprawić stosując:

- *Korektę na ciągłość:* każdej liczbie całkowitej *y* przypisujemy p-stwo odcinka od *y-0.5* do *y+0.5*, obliczone z odpowiedniego rozkładu normalnego.

Możemy stosować np. do rozkładu dwumianowego.

Przykład (powtórzony, z korektą na ciągłość)

- Załóżmy, że Y ma rozkład dwumianowy z n=40 i p=0.25.
- Obliczyć z korektą na ciągłość p-stwo, że Y jest pomiędzy 10 i 15 (włącznie).
- z_1= z_2=

P(10≤Y≤15)≈

[Wynik dokładny = 0.534.]

Przykład: wyniki testu.

Załóżmy, że μ = 100 i σ = 16, rozkład (ocen studenckich) ma kształt dzwonu (normalny).

Jakie jest p-stwo, że losowo wybrany student uzyska wynik między 120 a 140 punktów?

Rozkłady próbkowe: notacja

- Rozważmy populację o pewnym rozkładzie, np.:
  - *normalnym* N(μ, σ), lub
  - *dwupunktowym,* np. P(Y=sukces)=p, P(Y=porażka)=1-p
- Parametry populacji: μ i σ, lub p.
- Bierzemy próbę o rozmiarze **n** z populacji. Wynik:
  - $y_1, \ldots y_n$, lub
  - y = sumaryczna liczba sukcesów.
- Obliczamy estymatory
  - $\overline{y}$ i s, lub
  - $\hat{p}$
- Gdy **n** jest duże, estymatory są na ogół bliskie parametrom.

Rozkłady próbkowe, cd.

- Jak bardzo estymatory mogą sią różnić od prawdziwych parametrów?
- Co się stanie, jeżeli wylosujemy inną próbę?
- Otrzymamy inne wartości $\overline{y}$, s, $\hat{p}$ ....

Meta-eksperyment

- Wyobraźmy sobie, że powtarzamy eksperyment wiele razy.
- Interesuje nas rozkład wszystkich możliwych do uzyskania wartości $\overline{y}$, s lub $\hat{p}$.
- Taki rozkład będziemy nazywali rozkładem próbkowym estymatora.
- Zwykle próbkujemy tylko raz, ale rozkłady próbkowe można obliczyć teoretycznie.

Rozkład próbkowy estymatora $\hat{p}$ dla parametru *p* w rozkładzie dwupunktowym

- Y = liczba sukcesów w n próbach
- y = zaobserwowana liczba sukcesów
- $\hat{p}$ = Y/n jest estymatorem p
- Zasada: rozkład $\hat{p}$ sprowadza się do rozkładu Y (zob. przykłady).

Przykład:

- Producent ocenia, że 2% jego wyrobów jest wadliwych. Wyroby te paczkuje się po 40.
- Y = liczba wadliwych wyrobów w losowo wybranej paczce. Y ma rozkład .....
- Niech $\hat{p}$ = Y/40 = frakcja elementów wadliwych.
- P($\hat{p}$ = k/40) =

Rozkład $\hat{p}$ wyznaczamy przy pomocy (dwumianowego) rozkładu Y!

$$\Pr(\hat{p}=0)=\Pr(Y=0)=(1)(.02)^0(.98)^{40}=0.45$$
$$\Pr(\hat{p}=0.025)=\Pr(Y=1)=(40)(.02)^1(.98)^{39}=0.36$$
$$\Pr(\hat{p}=0.05)=\Pr(Y=2)=(780)(.02)^2(.98)^{38}=0.14$$
$$\Pr(\hat{p}=0.075)=\Pr(Y=3)=(9880)(.02)^3(.98)^{37}=0.04$$
$$\Pr(\hat{p}\geq 0.1)=\Pr(Y\geq 4)\approx 0.01$$

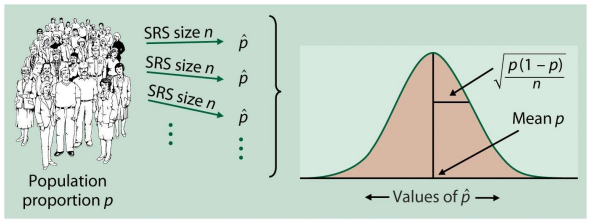
- Gdybyśmy otworzyli tysiące paczek, to rozkład frakcji liczby wadliwych elementów w paczce byłby dalece zgodny z rozkładem wyliczonym teoretycznie.
- Prawdziwa wartość **p** wynosi 0.02, i nie jest nawet możliwa do zaobserwowania w pojedynczym eksperymencie, ale na ogół $\hat{p}$ jest bliskie 0.02:
  - P-stwo tego, że $\hat{p}$ = 0.025 wynosi 36%.
  - P-stwo tego, że estymator będzie różnić się nie więcej niż o 0.03 od prawdziwej wartości to ...
- Zatem, jeżeli znajdziemy 3 lub więcej wyrobów wadliwych w jednej paczce mamy podstawy, żeby kwestionować twierdzenie producenta o **p**!

Przykład (cd., n=40 i p = 0.02.):

Jakie jest p-stwo, że estymator częstości przekracza dwukrotnie (lub więcej) prawdziwą wartość?

Zależność od rozmiaru próby:

- Y ma rozkład B(n,p)
- μ_Y=np
- Var (Y)=np(1-p)
- $\hat{p}$ =
- $\mu_{\hat{p}}$ =
- Var ($\hat{p}$)=

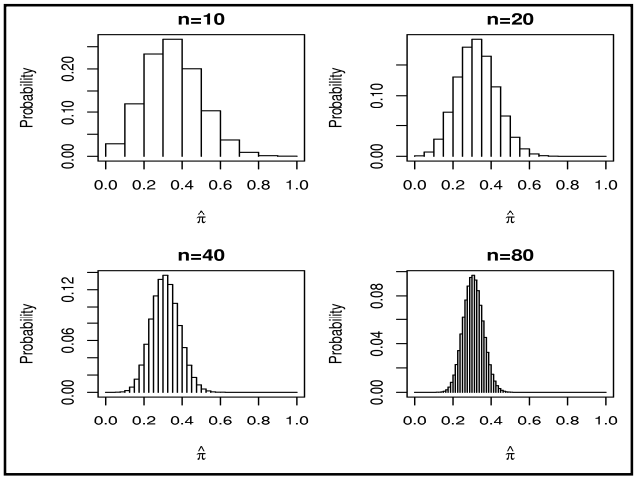


Wnioski:

- Gdy n rośnie, to wariancja .....

i estymator .....

- Na kolejnym rysunku p=0.3, n=10, 20, 40, 80.



Rozkład $\hat{p}$ (gdy p=0.3)

n	P(0.25≤ $\hat{p}$ ≤0.35)
10	0.5
20	0.535
40	0.612
80	0.728
500	0.987

Rozkład próbkowy dla *średniej* z *rozkładu normalnego*

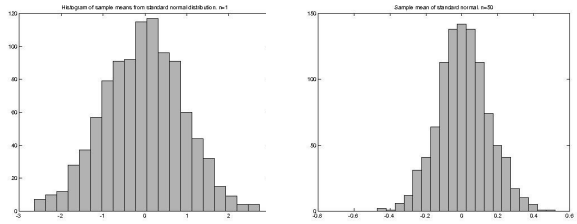
- Dana jest próba o rozmiarze n z populacji normalnej.
- Obserwujemy średnią próbkową.
- Jak daleko od μ może być $\bar{y}$?
- Odpowiemy na to pytanie znajdując rozkład próbkowy $\bar{y}$.

- Wyobraźmy sobie wielokrotne próbkowania: za każdym razem liczymy $\bar{y}$ ...
- Możemy o tym myśleć jak o eksperymencie, w którym obserwacjami są średnie (meta-eksperyment).
- Jaki jest rozkład tych średnich??

- FAKT 1 : Suma dwóch zmiennych niezależnych o rozkładzie normalnym ma rozkład normalny.
- FAKT 2 : Jeżeli X ma rozkład normalny to Y=aX+b, gdzie a ≠ 0, ma również rozkład normalny.

Rozkład $\overline{Y}$:

Histogramy średnich z rozkładu standardowego normalnego
Rozmiary pojedynczych prób n=1 and n=50.
Liczba powtórzeń eksperymentu N=1000.



Przykład: Y ~ N(30, 6).

- Bierzemy 10 próbek o rozmiarze n = 9:

$\overline{y}$	29.89	32.27	31.19	30.86	28.68
s	5.74	5.01	6.06	6.25	6.31

$\overline{y}$	29.60	30.02	31.19	29.84	30.27
s	6.83	3.81	5.13	4.82	4.90

Sprawdzenie:

- Rozkład $\overline{Y}$ ma wartość oczekiwaną=…..

i odchylenie standardowe SD=…..

- Oczekujemy, że średnia próbkowa:
  - Z p-stwem 0.95 będzie w odległości nie większej niż 1.96 SD od μ, tzn. pomiędzy ..... a .....
  - Z p-stwem 0.8 będzie w odległości nie większej niż 1.28 SD od μ, tzn. pomiędzy 27.4 a 32.6.
  - … 0.68 … 1 SD … tzn. pomiędzy 28 a 32.

Podsumowanie:

- Jeżeli Y ma rozkład normalny to rozkład (próbkowy) $\overline{Y}$ jest normalny.
- Niezależnie od rozkładu Y wartość oczekiwana średniej próbkowej= ……., a odchylenie standardowe średniej=…….

Centralne Twierdzenie Graniczne:

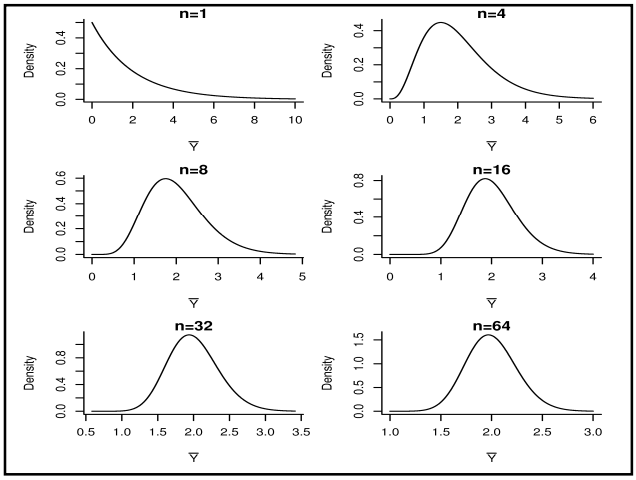
Nawet jeżeli *Y* nie ma rozkładu normalnego, to dla dużych *n* rozkład $\overline{Y}$ jest bliski normalnemu, gdy rozkład *Y* ma skończoną wariancję.

Szybkość zbieżności w CTG:

- Co to znaczy ``duże'' n ?

➢ Jeżeli Y ma rozkład normalny to wystarczy n= ...

➢ Jeżeli rozkład Y jest w przybliżeniu symetryczny i nie ma „ciężkich ogonów", to n=30 jest „dość duże".



Uwagi:

- Wiele metod statystycznych używa przybliżeń rozkładem normalnym.
- Techniki prezentowane dalej na tym kursie (np. przedziały ufności i testowanie hipotez dla μ) mogą być stosowane także wtedy, gdy rozkład obserwacji nie jest normalny, ale rozmiar prób jest „dość duży” i rozkład nie ma „ciężkich ogonów”.
- Założenie o niezależności poszczególnych obserwacji jest niezbędne.

Problem: Czas pracy lampy jest wykładniczy z wartością oczekiwaną 10000h. Oszacuj prawdopodobieństwo, że łączny czas pracy 20-tu lamp będzie krótszy niż 180kh.

Zadanie „dwustopniowe”:

Niech X będzie liczbą tych dużych prób spośród 10-ciu, dla których średnie różnią się od μ o więcej niż $\sigma/\sqrt{n}$. Oblicz P(X=2).