Estymacja punktowa i przedziałowa

Marta Zalewska
Zakład Profilaktyki Zagrożeń Środowiskowych i Alergologii

- Populacja
- Próba losowa (próbka)
- Parametry rozkładu
- Estymatory (statystyki)
- Własności estymatorów
- Błąd estymacji
- Przedziały ufności

Populacja (zbiorowość)

- Rozważamy zbiór elementów podlegających badaniu, ze względu na jedną cechę (na razie).

Badanie kompletne (całkowite, spis)

- Przebadane są wszystkie elementy zbioru (populacji).
- Dostarcza pełnej informacji o badanej cesze populacji.
- Często takie badanie jest niecelowe, kosztowne, czasochłonne bądź niewykonalne.
- Badaniami kompletnymi statystyka matematyczna nie zajmuje się.

Badanie reprezentacyjne

- Polega na wylosowaniu pewnej grupy przedstawicieli licznej populacji.
- Powiedzmy, że wylosowano 20 noworodków w celu poznania cech fizycznych dzieci urodzonych w Warszawie w tym roku.
- Przypuśćmy, że interesującą nas cechą jest zmienna losowa X =„ciężar ciała noworodka losowo wybranego z populacji”
- Dysponujemy ciągiem 20 liczb (w kg), możemy narysować dystrybuantę empiryczną.
- Rozkład badanej cechy w populacji utożsamiamy z rozkładem prawdopodobieństwa fikcyjnej zmiennej losowej X.

Populacja i próbka losowa

- Badamy próbkę losową, aby dowiedzieć się czegoś o populacji (zbiorowości)
- Zakładamy że mamy do czynienia ze zmiennymi losowymi $X_1, X_2,\ldots X_n$ i dane są realizacje tych zmiennych losowych $x_i = X_i(\omega),\ldots$Nie znamy natomiast rozkładu prawdopodobieństwa, z którego te zmienne są wylosowane.
- Próbujemy dowiedzieć się czegoś o nieznanym rozkładzie prawdopodobieństwa tych zmiennych na podstawie obserwacji $x_1, x_2,\ldots, x_n$

Najczęściej zakładamy, że próbka jest tzw. prostą próbką losowa tzn:

1) każda jednostka populacji ma takie samo prawdopodobieństwo trafienia do próbki

2) każda kolejna jednostka jest wybierana do próbki niezależnie.

Są dwa podstawowe rodzaje losowania próbki:

1) Losowanie bez zwracania (zależne)
2) Losowanie ze zwracaniem (jednostka może wielokrotnie trafić do tej samej próbki, losowanie niezależne)

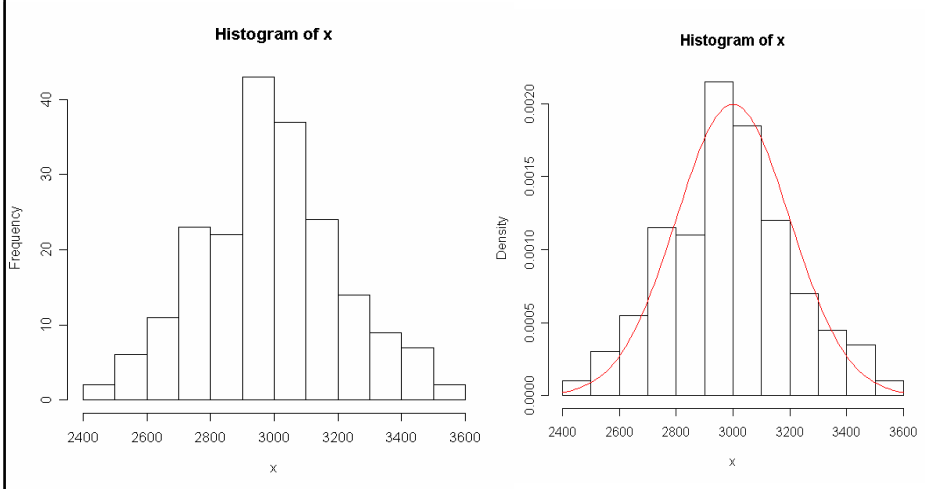
Częściej stosowane jest losowanie bez zwracania.

Jeśli populacja jest skończona to spełnienie warunku niezależności wymaga losowania ze zwracaniem. Jest to schemat matematycznie prostszy.
Dla dużej populacji praktycznie zaciera się różnica pomiędzy obydwoma sposobami losowania.

Przykład. Analiza cen komputerów

Populacja: wszystkie sklepy komputerowe w Polsce
Cecha: cena komputera (traktujemy jako cechę ciągłą)
Ceny odkładamy na osi poziomej, na osi pionowej odkładamy ile razy konkretna cena się powtórzyła, wyrażoną w procentach.
Otrzymujemy rozkład wartości ceny komputerów w Polsce (prawdopodobnie krzywą dzwonową) – pole pod tą krzywą równe jest 1. Pole zakreślone między dwoma cenami – przedstawia % sklepów, w których ceny znajdują się w tym przedziale. Jest to **rozkład cechy w populacji**.
Najdroższe sklepy będą po prawej stronie, najtańsze po lewej stronie osi poziomej.

Wyniki próbki 200 elementowej



Rozkład cechy w populacji traktujemy jako rozkład prawdopodobieństwa zmiennej losowej X (oznaczającej wartość cechy dla jednostki losowo wybranej z populacji).

Rozkład prawdopodobieństwa to jest **charakterystyka populacji.**

Parametry rozkładu prawdopodobieństwa np. **$E(X)=\mu$, $Var(X)=\sigma^2$ (na ogół nieznane)** traktujemy jako skrótowe charakterystyki populacji

Wartość oczekiwana jest charakterystyką populacji sklepów - jest średnią ceną ze wszystkich sklepów.

Odchylenie standardowe mówi jak średnio odchylają się wartości w pojedynczych sklepach od średniej.

Oba parametry są nieznane – aby je poznać należałoby zbadać wszystkie sklepy.

Zwykle dostępna jest tylko **próbka.** W naszym przypadku będzie to próba 200 sklepów.
Z punktu widzenia statystyki próbka - to **niezależne zmienne losowe** $X_1, X_2,\ldots X_{200}$ o takim samym rozkładzie prawdopodobieństwa jak X
X jest wzorcową zmienną – ceną komputera w losowo wybranym sklepie
$X_1, X_2,\ldots X_{200}$ są to ceny w 200 niezależnie wybranych sklepach.

Na podstawie próbki oblicza się próbkowe odpowiedniki wielkości populacyjnych.

Odpowiednikiem wartości oczekiwanej jest średnia (w przykładzie z 200 wartości) i jest nazywana estymatorem nieznanej liczby μ (mi), a wariancja z próbki jest estymatorem wariancji σ^2

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum X_i$$

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2$$

Estymatory to wielkości obliczone na podstawie próbki, które oszacowują nieznane parametry populacji.

Wyniki oszacowania ceny w losowo wybranych 200 sklepach:

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum X_i = 3001.4$$

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2 = 43126.41$$

Oszacowanie odchylenia standardowego wynosi: 207.67

Należy odróżnić estymator od wielkości estymowanej.
Estymatory to zmienne losowe, bo jeśli dane są losowe to wszystko, co policzymy na podstawie tych danych, też będzie losowe.

Przypuśćmy, że powtarzamy doświadczenie 10 razy, tzn. 10 razy losujemy 200 sklepów
z tej samej populacji. I otrzymujemy : 10 nowych średnich

2987.8 2997.4 2987.4 3002.3 2989.1 3034.2
3000.9 3017.2 2998.9 2987.6

Podstawowe statystyki:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2987	2988	2998	3000	3002	3034

Liczymy wartość oczekiwaną i wariancję średniej.
Jaka jest wartość oczekiwana

$$E(\overline{X}) = \mu$$

Jaka jest wariancja?

$$Var(\overline{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) =$$

Bo są niezależne

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

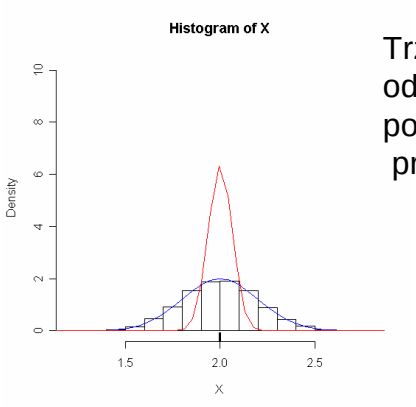
$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

A odchylenie standardowe ?

$$D(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

Przykład

Zmienna losowa X~N(2,0.2^2) na rysunku kolor niebieski
Zmienna losowa $\overline{X}$ ~N(2, $(0.2/\text{sqrt}(10))^2$)
n=10 kolor czerwony, odchylenie standardowe =0.063



Trzeba podzielić odchylenie standardowe pojedynczej obserwacji przez sqrt(n)

σ - Odchylenie standardowe w populacji (rozrzut cen we wszystkich sklepach)

$\hat{\sigma}$ - Estymator odchylenia standardowego w populacji (rozrzut cen oszacowany na podstawie próbki)

$\frac{\sigma}{\sqrt{n}}$ - Błąd standardowy średniej próbkowej (na ile ona odchyla się średnio od średniej populacyjnej)

$\frac{\hat{\sigma}}{\sqrt{n}}$ - Estymator błędu standardowego średniej próbkowej (oszacowanie dokładności z jaką estymujemy średnią populacyjną)

Parametr θ

- Odgrywa rolę identyfikatora rozkładu prawdopodobieństwa

Przykład.

- Liczba wypadków drogowych w ciągu tygodnia ma w przybliżeniu rozkład **Poissona** z parametrem $\boldsymbol{\theta} = \lambda$ Niech liczby $X_1, X_2, \ldots X_n$- oznaczają liczby wypadków w kolejnych tygodniach .
- Zbiór możliwych wartości $\boldsymbol{\theta}$ - przedział nieograniczony od 0 do nieskończoności

$\boldsymbol{\theta}$ jest zarówno wartością oczekiwaną, jak i wariancją zmiennej losowej X opisującej liczbę wypadków w ciągu tygodnia.

Estymacja

Estymacja - szacowanie parametrów populacji na podstawie obserwacji uzyskanych w próbie losowej

θ - **theta** jest parametrem rozkładu cechy X w populacji

(theta może być liczbą, parą liczb, itp.)

Nieznaną wartość θ szacujemy na podstawie n - elementowej próbki losowej $(x_1, x_2, \ldots x_n)$

Estymator (punktowy) jest funkcja próby przybliżającą wartość parametru theta .

Rozkład próbkowy estymatora

Rozkład próbkowy estymatora jest rozkładem prawdopodobiestwa wszystkich możliwych wartości, który może przyjąć estymator, kiedy jest obliczany z próbek losowych tego samego rozmiaru, wylosowanych z tej samej populacji.

Aby zobaczć losowość, trzeba wyobrazić sobie dużo prób, za każdym razem ekperyment losowy z nową próbką o tej samej liczebności. Stąd się bierze rozkład próbkowy. Np. z populacji 100 elementowej można wybrać

$$\binom{100}{10} \approx 17000000000000$$

różnych próbek 10-elementowych!

Estymacja

Populacja	Próbka losowa
Opisywana przez parametry ustalone ale nieznane	Estymatory obliczane z próbki znane i losowe
1. średnia populacyjna	średnia próbkowa
2. wariancja populacyjna	wariancja próbkowa
3. odch. standandardowe w populacji	odch. standardowe w próbce
4. wsp. zmienności w populacji	wsp. zmienności w próbce
5. mediana populacyjna	mediana z próbki
6. wsk. struktury w populacji	wsk. struktury w próbce

Przykład:
Badamy populację o rozkładzie z wartością oczekiwaną E(X)
Średnia arytmetyczna

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}$$

z n - elementowej próby losowej jest **nieobciążonym estymatorem** wartości oczekiwanej populacji

Przykład:

- **nieobciążony estymator** wariancji populacji (bez falki)

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$$

- **obciążony estymator** wariancji populacji –wariancja próbkowa

$$\tilde{S}^2 = \frac{1}{n}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$$

Najbardziej naturalny estymator wariancji jest obciążony

$$\hat{\sigma}^2 = \tilde{S}^2$$

JEST OBCIĄŻONYM ESTYMATOREM WARIANCJI POPULACJI

$$\sigma^2 = Var(X)$$

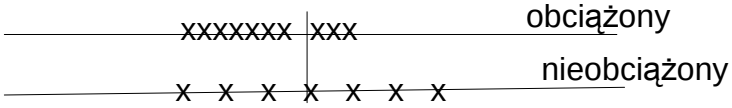
Aby estymator wariancji był nieobciążony, będziemy dzielić sumę kwadratów odchyleń przez 1 / n-1

Aby estymator był **nieobciążony** to jego wartość oczekiwana musi być równa estymowanemu parametrowi populacji

$$E(S^2) = E(\frac{n}{n-1}\tilde{S}^2) = \sigma^2$$

Estymator wariancji ma swoją wartość oczekiwaną, ma swoje odchylenie standardowe i wariancję

$$Var(\tilde{S}^2) < Var(S^2)$$



Przykład

Czasy wykonania pewnej analizy wyniosły:

14.1, 15.1, 13.8, 16.4, 13, 15.2, 14.8, 16.4, 16.1, 15.1

Zbudować estymatory nieznanych parametrów populacji na podstawie próbki.

Jaka jest interpretacja czasów wykonania analizy
w naszym przykładzie.
Jeżeli weźmiemy typową zmienna losową opisującą
czas wykonania analizy to :

$$E(X) = \mu$$
$$Var(X) = \sigma^2$$
$$D(X) = \sigma$$

Są to nieznane parametry

μ Jest to średni czas dla wszystkich

σ Jest to średni rozrzut dookoła średniej

Nieobciążony estymator wartości oczekiwanej populacji:

$$\hat{\mu} = \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad =15$$

Nieobciążony estymator wariancji populacji:

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2] \quad = 0.148$$

Nieznany parametr populacji	**Estymator**	**Własności estymat.**
średnia μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	nieobc. zgodny
wariancja σ^2 (μ znane)	$S_1^2 = \frac{1}{n}\sum(X_i - \mu)^2$	nieobc. zgodny
wariancja σ^2 (μ nieznane)	$S_*^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$	obciążony zgodny
wariancja σ^2 (μ nieznane)	$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$	nieobc. zgodny
odch. stand. σ	$S = \sqrt{\frac{1}{n-1}\sum(X_i - \bar{X})^2}$	obciążony zgodny
wsp. zmienności $\nu = \frac{\sigma}{\mu}$	$V = S/\bar{X}$	obciążony zgodny
mediana popul.	$med(X_1, \ldots, X_n)$	obciąż. *) zgodny
wskaźnik strukt. p	$\widehat{p} = \frac{k}{n}$	nieobc. zgodny

*) jeśli rozkład w populacji jest symetryczny, to nieobciążony.

wartość estymatora	—	wartość parametru populacji	=	błąd estymacji
Np. $\bar{X}$	—	μ	=	błąd estymacji
Np. S^2	—	σ^2	=	błąd estymacji

Jakość estymatora mierzy się błędem średniokwadratowym lub błędem standardowym.

Błąd średniokwadratowy jest to średni kwadrat błędu (uśredniamy ze względu na rozkład próbkowy estymatora). $BŚK = E[(\hat{\theta} - \theta)2]$

błąd standardowy $=$ $\sqrt{\text{błąd średniokwadratowy}}$

(Błąd standardowy estymatora $\bar{X}$) $= \frac{\sigma}{\sqrt{n}}$

(Estymator błędu stand. estymatora $\bar{X}$) $= \frac{S}{\sqrt{n}}$

Estymacja przedziałowa

Pamiętajmy dane w statystyce traktujemy jako zmienne losowe.

Przyjmujemy założenie o tym, jaki jest rozkład prawdopodobieństwa.

Obliczamy estymatory nieznanych parametrów populacji.

Estymatory – oszacowania nieznanych parametrów populacji obliczamy na podstawie próbki.

Estymacja przedziałowa – chcemy, aby nieznany parametr znalazł się między dwoma oszacowaniami z góry określonym prawdopodobieństwie

Przedziały ufności

DEFINICJA. Niech θ będzie nieznanym parametrem populacji a $X_1, \ldots, X_n$ – wylosowaną próbką. Mówimy, że $[\hat{\theta}_1, \hat{\theta}_2]$ jest **przedziałem ufności** dla θ na poziomie $1-\alpha$, jeśli $\hat{\theta}_1 = \hat{\theta}_1(X_1, \ldots, X_n)$ i $\hat{\theta}_2 = \hat{\theta}_2(X_1, \ldots, X_n)$ oraz

$$\mathbb{P}\left(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2\right) \geq 1 - \alpha.$$

Zauważmy, że przedział ufności ma końce, które są zmiennymi losowymi

Przedział ufności

- Jest obliczony na podstawie danych
- Jest założone prawdopodobieństwo, że przedział ufności zawiera nieznany param populacji. Pamiętajmy, że końce przedziału są losowe a parametr jest nielosowy.
- Poziom ufności – przeważnie 95% jest to prawdopodobieństwo, że przedział zawiera estymowany parametr populacji (może być: 99%, 99,9%, 90%)

Przedział ufności c.d.

- Przedział na poziomie ufności 0.95 to taki przedział, że jak wiele razy będziemy powtarzali eksperyment, to średnio 95% wyznaczonych w ten sposób przedziałów zawiera szacowany parametr, a około 5% nie zawiera ich. Oczywiście nigdy nie wiemy, czy trafimy na taki przedział, który zawiera szacowaną wartość czy też nie. Dlatego mówimy, że z ufnością 0.95 (lub 95%) jesteśmy pewni, że w danym przedziale zawiera się szacowany parametr.

Tworząc przedział dla nieznanego parametru theta

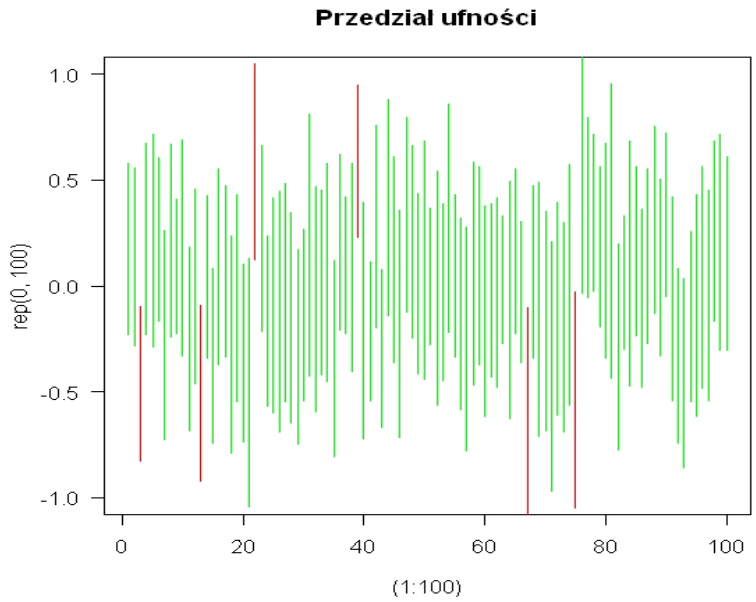
$$P(\hat{\theta}_1 \le \theta \le \hat{\theta}_2) \ge 1-\alpha$$

Konstruujemy dwa estymatory: $\hat{\theta}_1$ i $\hat{\theta}_2$

które dają się policzyć na podstawie danych z próbki.
Chcemy, aby z dużym prawdopodobieństwem nieznany parametr znalazł się w tym przedziale.
W przykładzie skonstruowaliśmy estymator $\hat{\mu} =$ 15

A teraz chcemy $P(\hat{\mu}_1 \le \mu \le \hat{\mu}_2) \ge 0.95$

$1-\alpha$ to poziom ufności



$$P(\hat{\mu}_1 \le \mu \le \hat{\mu}_2) \ge 0.95$$

$$N(\mu, \sigma^2)$$

σ^2 Znana wariancja w populacji

$$[\bar{x} - \frac{z\,\sigma}{\sqrt{n}}, \bar{x} + \frac{z\,\sigma}{\sqrt{n}}]$$

$$\frac{\sigma}{\sqrt{n}}$$

$\bar{x}$

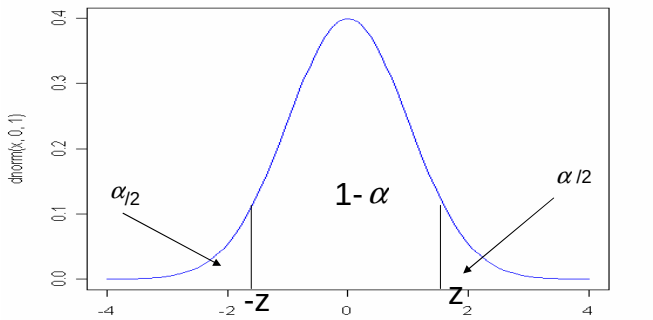
z = 1.96 kwantyl rozkładu $N(0,1^2)$

1. Przedział ufności dla μ Próbka z rozkładu $N(\mu,\sigma^2)$

Należy wziąć przedział: σ^2 znane

$$[\bar{x} - \frac{z\,\sigma}{\sqrt{n}}, \bar{x} + \frac{z\,\sigma}{\sqrt{n}}]$$

Na lewo od z jest pole 1 - α + α /2 z= kwantyl rzędu(1 - α /2)



„z" =1.96 =kwantyl rozkładu normalnego rzędu(1-0.05/2) pomiędzy z i –z jest pole 1 - α

Tzn. (pole na lewo od 1.96)=0.975

Przykład: Skonstruować przedział ufności dla μ
na poziomie 95% jeżeli wiemy, że $\bar{x}$ = 15 i σ =1

Jak znaleźć kwantyl 1- α=0,95
To α =0.05 Ile jest 1- α /2 ? =0.975

z=$z_{0.975}$=1.96

$$[\bar{x} - \frac{z\,\sigma}{\sqrt{n}}, \bar{x} + \frac{z\,\sigma}{\sqrt{n}}]$$

d/2
0.619795

$$[15 - \frac{1.96 \cdot 1}{\sqrt{10}}, 15 + \frac{1.96 \cdot 1}{\sqrt{10}}]$$

[14.38020 , 15.61980]

Mówimy: Moje oszacowanie średniego czasu wykonania analizy wskazuje, że ten czas powinien się mieścić w przedziale [14.38020 ,15.61980]
Zaufanie do tego wniosku wynosi 95%

W przybliżeniu:

95% przedział ufności

Parametr populacji (mi) =

(średnia próbkowa $\pm$ 2 * (błąd standardowy średniej)

$$\bar{x} \pm 2 * \frac{\sigma}{\sqrt{n}}$$

$$\mu = 15 \pm 0.63$$

Na poziomie ufności 0.95

Zadanie.
Z tych samych danych skonstruować przedział ufności na poziomie 99%

$$\alpha = 0.01$$

$$1 - \frac{\alpha}{2} = 1 - \frac{0.01}{2} = 1 - 0.005 = 0.995$$

$$z = z_{0.995} = 2.58$$

$$[\bar{x} - \frac{z\sigma}{\sqrt{n}}, \bar{x} + \frac{z\sigma}{\sqrt{n}}]$$

$$[15 - \frac{2.58 \cdot 1}{\sqrt{10}}, 15 + \frac{2.58 \cdot 1}{10}]$$

$\mu = 15 \pm 0.8145487$

[14.18545, 15.81455]

Na poziomie ufności 0.99

Rozkład t lub rozkład t-Studenta)

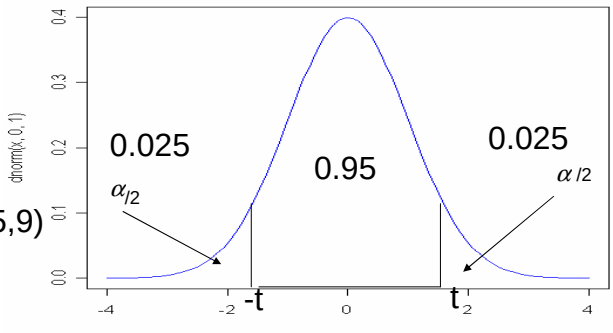
Dysponujemy wynikami **n** pomiarów, dla których możemy wyznaczyć estymatory parametrów populacyjnych, jak średnia i odchylenie standardowe S lub wariancja S^2**,** nie znamy natomiast odchylenia standardowego w populacji. Zagadnienie to rozwiązał (w 1908r.) W.S.Gosset (pseudonim Student) podając funkcję zależną od tzw. stopni swobody (df) i poziomu istotności α

Wartości krytyczne $t = t(\alpha, n-1)$ rozkładu t-Studenta z tablicy
Stopnie swobody związane są z licznością próbki df= n-1

2. Przedział ufności dla μ Próbka z rozkładu $N(\mu, \sigma^2)$
Nie znamy wariancji σ^2

Należy wziąć przedział:

$$[\overline{X} - t(\alpha; n-1)\frac{S}{\sqrt{n}}, \overline{X} + t(\alpha; n-1)\frac{S}{\sqrt{n}}]$$



Dla t(0.05,9)
t=2.2622

$t = t(\alpha, n-1)$ „t" wartość krytyczna rozkładu t Studenta z n-1 stopniami swobody
„S" jest estymatorem σ (t tak jak z tylko dla innego rozkładu)

Wartości krytyczne $t(\alpha; n-1)$ rozkładu t – Studenta są stablicowane.
Stopnie swobody (n-1) w tablicy oznaczone „r” znajdujemy w odpowiednim wierszu , a zadane α w odpowiedniej kolumnie.
Na przecięciu wiersza i kolumny odczytujemy wartość t , dla n-1= 9 i α =0.05 t= 2.2622

α

Dla rozkładu t tablicuje się sumę dwóch ogonów
Nie tak, jak dla rozkładu normalnego.

Jeżeli chcemy mieć przedział jednostronny to aby mieć poziom 95% odczytujemy w tablicach t Studenta dla 2 α czyli dla 0.10.

http://www.math.uni.wroc.pl/~zpalma/tablicetstudenta.pdf

Przykład.
Wykorzystamy dane z poprzedniego przykładu:
Obliczone na podstawie próbki:
Średnia=15

Wariancja=1.275556
Odchylenie_stand=1.129405

$$[\overline{X} - t(\alpha; n-1)\frac{S}{\sqrt{n}}, \overline{X} + t(\alpha; n-1)\frac{S}{\sqrt{n}}]$$

$$[15 - 2.2622\frac{1.129405}{\sqrt{10}}, 15 + 2.2622\frac{1.129405}{\sqrt{10}}]$$

$\mu \in$ [14.1920, 15.8079] na poziomie ufności
Zaufanie do tego wniosku wynosi 95%

Długość przedziału:

$$d = 2t(\alpha; n-1)\frac{S}{\sqrt{n}}$$

Przedziały jednostronne:

$$(-\infty, \overline{X} + t(2\alpha; n-1)\frac{S}{\sqrt{n}})$$

$$(\overline{X} - t(2\alpha; n-1)\frac{S}{\sqrt{n}}, +\infty)$$

Przykład.
Oszacować przeciętną ilość punktów uzyskiwanych na klasówce mając następujące dane:

n=300, $\sum x_i = 176.566$ $\sum x_i^{\,2} = 107.845$

Populacja:
Słuchacze kursu statystyki
Cecha X:
Ilość punktów zdobyta na klasówce
Założenie:
Cecha X ma rozkład normalny $N(\mu,\sigma^2)$
Zadanie:
Oszacować parametr
Technika statystyczna:
Przedział ufności dla średniej μ
Poziom ufności $1-\alpha$ =0.95

Obliczenia:

$\bar{x} = \frac{\sum x_i}{n}$ =176.566 / 300 = 0.589

$S^2 = \frac{1}{n-1}\left[x_1^2 + x_2^2 + \cdots x_n^2\right] - \bar{x}^2$ $\qquad S^2 = \frac{\sum x_i^2}{n-1} - \bar{x}^2$

$S^2 = \frac{107.845}{299} - (0.589)^2$ =0.013

$S = \sqrt{S^2} = \sqrt{0.013} = 0.114$

$t(\alpha; n-1)$ = t (0.05,299) =ok. 1.96 jak dla rozkł norm.

$t(0.05,9)\frac{S}{\sqrt{n}}$ = 1.96* 0.114/sqrt(300)= 0.0129

(0.589-0.013,0.589+0.013)

Odpowiedź: $\mu \in$ (0.576,0.602) z zaufaniem 95%

Przybliżony przedział ufności dla wskaźnika struktury

$$\left[\hat{p} - \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} * z,\ \hat{p} + \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} * z\right]$$

Dla poziomu istotności $1-\alpha$ = 0.95

$z = z_{1-\frac{\alpha}{2}} = z_{0.975}$ =1.96 Z kwantyl rozkładu $N(0,1^2)$

Uwaga - n musi być duże

Przykład.

Z populacji wyborców pobrano próbkę 1000 osób i okazało się, że wśród nich 300 popiera partię X. Podać przedział ufności dla frakcji wyborców popierających partię X w populacji na poziomie ufności (1-0.05)=95%.

Populacja:
Wyborcy
Cecha X:
Poparcie dla partii X
Założenie:
Cecha X ma rozkład D(p)=Bin(1,p)
Zadanie: oszacować parametr p
Technika statystyczna: przybliżony przedział ufności dla prawdopodobieństwa Poziom ufności 0.95

Przykład cd.

Obliczenia:

k=300
n=1000
m=20 mln
p?

$\hat{p}$ = k/n = 300/1000=0.3

$P(\hat{p}_1 \le p \le \hat{p}_2) \ge 0.95$ 1-0.025= 0.975

α = 0.05 $Z_{0.975}$=1.96

p: $\hat{p} \pm \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \cdot z$ = $0.3 \pm 1.96 \frac{\sqrt{0.3 \cdot 0.7}}{\sqrt{1000}}$ = $0.3 \pm$ 0.0284

p: [0.2716, 0.3284] Z ufnością 95%

Przeważnie przekazując badania opinii publicznej nie podaje się przedziału ufności lecz mówi się o błędzie (media podałyby: poparcie dla partii X wynosi 30%; błąd oszacowania $\pm$ 3%)