STATYSTYKA

Stanisław Jaworski

Katedra Ekonometrii i Informatyki
Zakład Statystyki

Wstęp

- Statystyka
- Populacja, próba, cecha

Wymagane wiadomości

- Dystrybuanta
- Wybrane rozkłady prawdopodobieństwa

Pytania

Estymacja punktowa

- Definicja
- Przykłady

Estymacja przedziałowa

- Definicje
- Jedna populacja
- Dwie populacje

Pytania

Badanie zależności między dwiema cechami losowymi

- Wprowadzenie
- Weryfikacja niezależności cech
- Opis ilościowy, estymacja

Badanie zależności między cechą losową a deterministyczną

- Model, opis ilościowy
- Weryfikacja niezależności cech
- Obszar ufności, Predykcja

Test Chi-kwadrat

- Wprowadzenie
- Oznaczenia
- Hipoteza oraz jej weryfikacja

Dynamika zjawisk

- Indeksy agregatowe

Analiza danych

- Szereg rozdzielczy
- Mierniki położenia i rozproszenia
- Koncentracja

Część I

Wprowadzenie

Statystyka

Nauka poświęcona metodom badania (analizowania) zjawisk masowych; polega na systematyzowaniu obserwowanych cech ilościowych i jakościowych oraz przedstawianiu wyników w postaci zestawień tabelarycznych, wykresów, itp.; posługuje się rachunkiem prawdopodobieństwa.

Statystyka matematyczna

Dział matematyki stosowanej oparty na rachunku prawdopodobieństwa; zajmuje się badaniem zbiorów na podstawie znajomości własności ich części.

- Populacja. Zbiór obiektów z wyróżnioną cechą (cechami).Obiektami mogą być przedmioty lub wartości cechy
- Próba. Wybrana część populacji podlegająca badaniu. Próba powinna stanowić reprezentację populacji w tym sensie, że częstości występowania w próbie każdej z badanych cech nie powinny się znacznie różnić od częstości występowania tych cech w populacji
- Cecha losowa. Wielkość losowa charakteryzująca obiekty danej populacji.

- Populacja. Zbiór obiektów z wyróżnioną cechą (cechami).Obiektami mogą być przedmioty lub wartości cechy
- Próba. Wybrana część populacji podlegająca badaniu. Próba powinna stanowić reprezentację populacji w tym sensie, że częstości występowania w próbie każdej z badanych cech nie powinny się znacznie różnić od częstości występowania tych cech w populacji
- Cecha losowa. Wielkość losowa charakteryzująca obiekty danej populacji.

- Populacja. Zbiór obiektów z wyróżnioną cechą (cechami).Obiektami mogą być przedmioty lub wartości cechy
- Próba. Wybrana część populacji podlegająca badaniu. Próba powinna stanowić reprezentację populacji w tym sensie, że częstości występowania w próbie każdej z badanych cech nie powinny się znacznie różnić od częstości występowania tych cech w populacji
- Cecha losowa. Wielkość losowa charakteryzująca obiekty danej populacji.

Rodzaje cech

- **Cecha niemierzalna** – zwana też jakościową – przyjmuje wartości nie będące liczbami (np. *kolor*, *płeć*,*smakowitość*)
- **Cecha mierzalna**– zwana też ilościową– przyjmuje pewne wartości liczbowe (np. *długość, wytrzymałość, ciężar*)
- **Cecha (mierzalna) skokowa** –zwana też dyskretną– nie przyjmuje wartości pośrednich (np. *ilość bakterii*, *ilość pracowników*, *ilość pasażerów*,).
- **Cecha (mierzalna) ciągła** przyjmuje wartości z pewnego przedziału liczbowego (np. *wzrost*, *waga*, *plon*,*czas obsługi*)

Rodzaje cech

- **Cecha niemierzalna** – zwana też jakościową – przyjmuje wartości nie będące liczbami (np. *kolor*, *płeć*,*smakowitość*)
- **Cecha mierzalna**– zwana też ilościową– przyjmuje pewne wartości liczbowe (np. *długość, wytrzymałość, ciężar*)
- **Cecha (mierzalna) skokowa** –zwana też dyskretną– nie przyjmuje wartości pośrednich (np. *ilość bakterii*, *ilość pracowników*, *ilość pasażerów*,).
- **Cecha (mierzalna) ciągła** przyjmuje wartości z pewnego przedziału liczbowego (np. *wzrost*, *waga*, *plon*,*czas obsługi*)

Rodzaje cech

- **Cecha niemierzalna** – zwana też jakościową – przyjmuje wartości nie będące liczbami (np. *kolor*, *płeć*,*smakowitość*)
- **Cecha mierzalna**– zwana też ilościową– przyjmuje pewne wartości liczbowe (np. *długość, wytrzymałość, ciężar*)
- **Cecha (mierzalna) skokowa** –zwana też dyskretną– nie przyjmuje wartości pośrednich (np. *ilość bakterii*, *ilość pracowników*, *ilość pasażerów*,).
- **Cecha (mierzalna) ciągła** przyjmuje wartości z pewnego przedziału liczbowego (np. *wzrost*, *waga*, *plon*,*czas obsługi*)

Rodzaje cech

- **Cecha niemierzalna** – zwana też jakościową – przyjmuje wartości nie będące liczbami (np. *kolor*, *płeć*,*smakowitość*)
- **Cecha mierzalna**– zwana też ilościową– przyjmuje pewne wartości liczbowe (np. *długość, wytrzymałość, ciężar*)
- **Cecha (mierzalna) skokowa** –zwana też dyskretną– nie przyjmuje wartości pośrednich (np. *ilość bakterii*, *ilość pracowników*, *ilość pasażerów*,).
- **Cecha (mierzalna) ciągła** przyjmuje wartości z pewnego przedziału liczbowego (np. *wzrost*, *waga*, *plon*,*czas obsługi*)

Dystrybuanta F jest funkcją określoną na zbiorze liczb rzeczywistych R wzorem

$$F(x) = P\{X \leq x\}, \qquad x \in R.$$

Najważniejsze własności dystrybuanty

1. $0 \leq F(x) \leq 1$
2. $F(-\infty) = 0, \ F(\infty) = 1$
3. dystrybuanta jest funkcją niemalejącą
4. $P\{a < X \leq b\} = F(b) - F(a)$

Zmienna losowa X ma **rozkład dwupunktowy** ($X \sim D(p)$), jeżeli z dodatnimi prawdopodobieństwami przyjmuje jedynie dwie wartości x_1 i x_2:

$$P(X = x_2) = p, \quad P(X = x_1) = 1 - p, \quad 0 < p < 1.$$

Zmienna losowa X ma **rozkład dwumianowy** ($X \sim B(n, p)$), jeżeli

$$P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, \ldots, n.$$

Schemat Bernoulliego. Wykonujemy dwuwynikowe doświadczenie. Wyniki nazywane są umownie *sukces* oraz *porażka*. Prawdopodobieństwo sukcesu wynosi p (porażki $1-p$). Doświadczenie wykonujemy w sposób niezależny n krotnie. Niech zmienną losową X będzie ilość sukcesów. Zmienna losowa X ma rozkład $B(n,p)$.

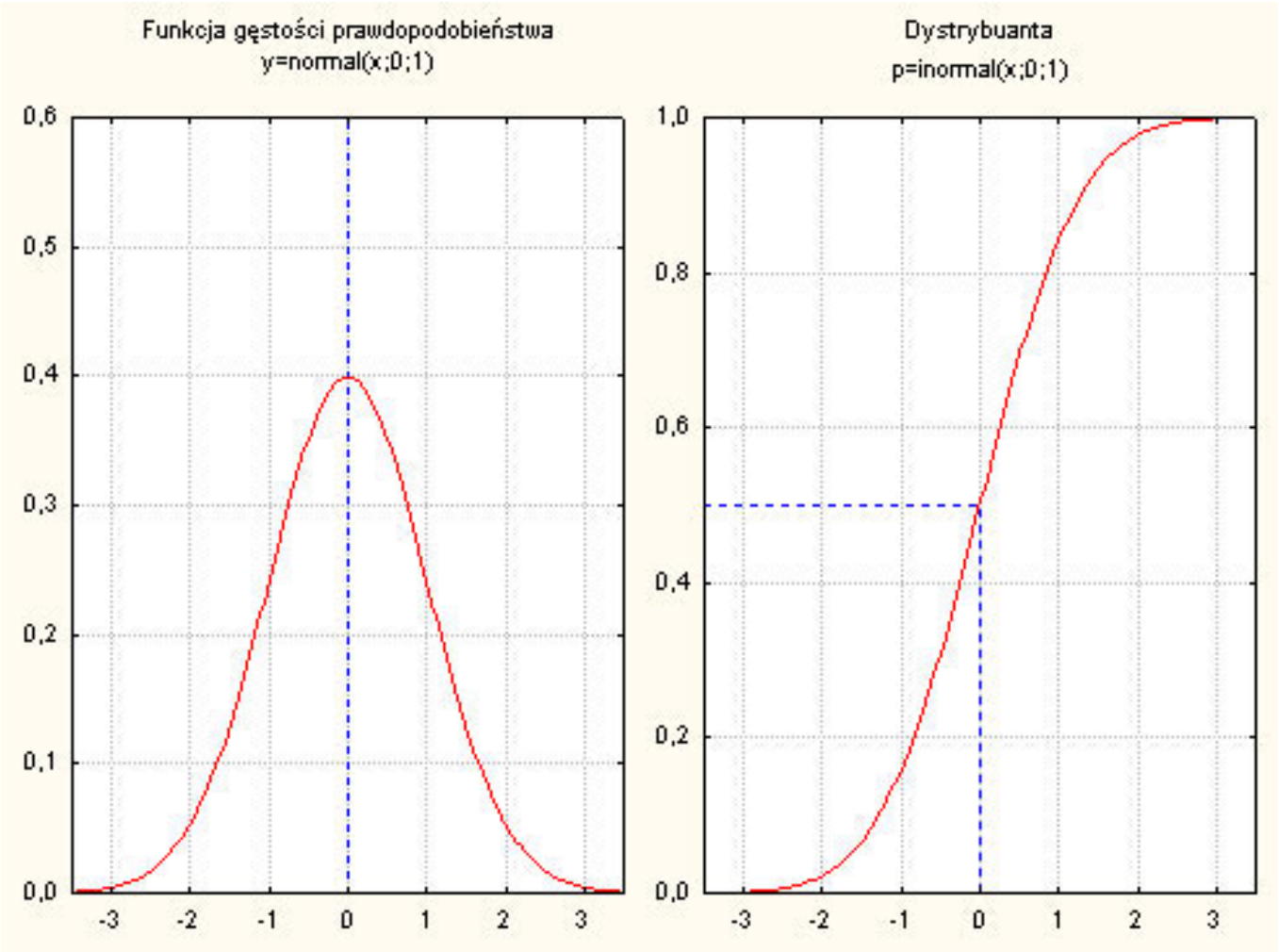
Rozkład normalny $N(\mu, \sigma^2)$. Zmienna losowa X ma rozkład normalny o wartości średniej μ i wariancji σ^2, jeżeli jej funkcja gęstości wyraża się wzorem

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

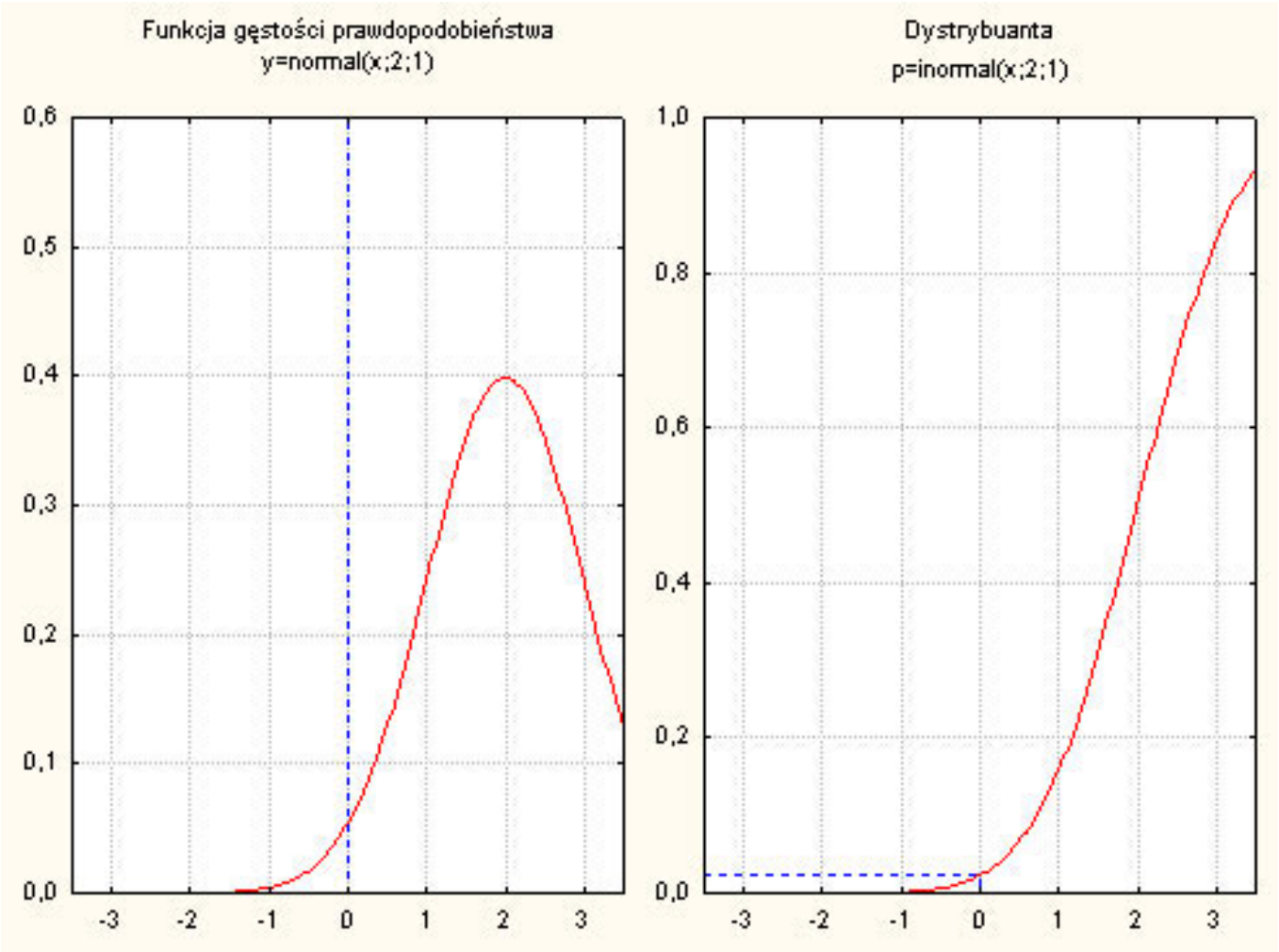
Prawo trzech sigm

$$P\{|X - \mu| < \sigma\} = 0.68268 \approx 0.68$$
$$P\{|X - \mu| < 2\sigma\} = 0.95550 \approx 0.96$$
$$P\{|X - \mu| < 3\sigma\} = 0.99730 \approx 0.997$$

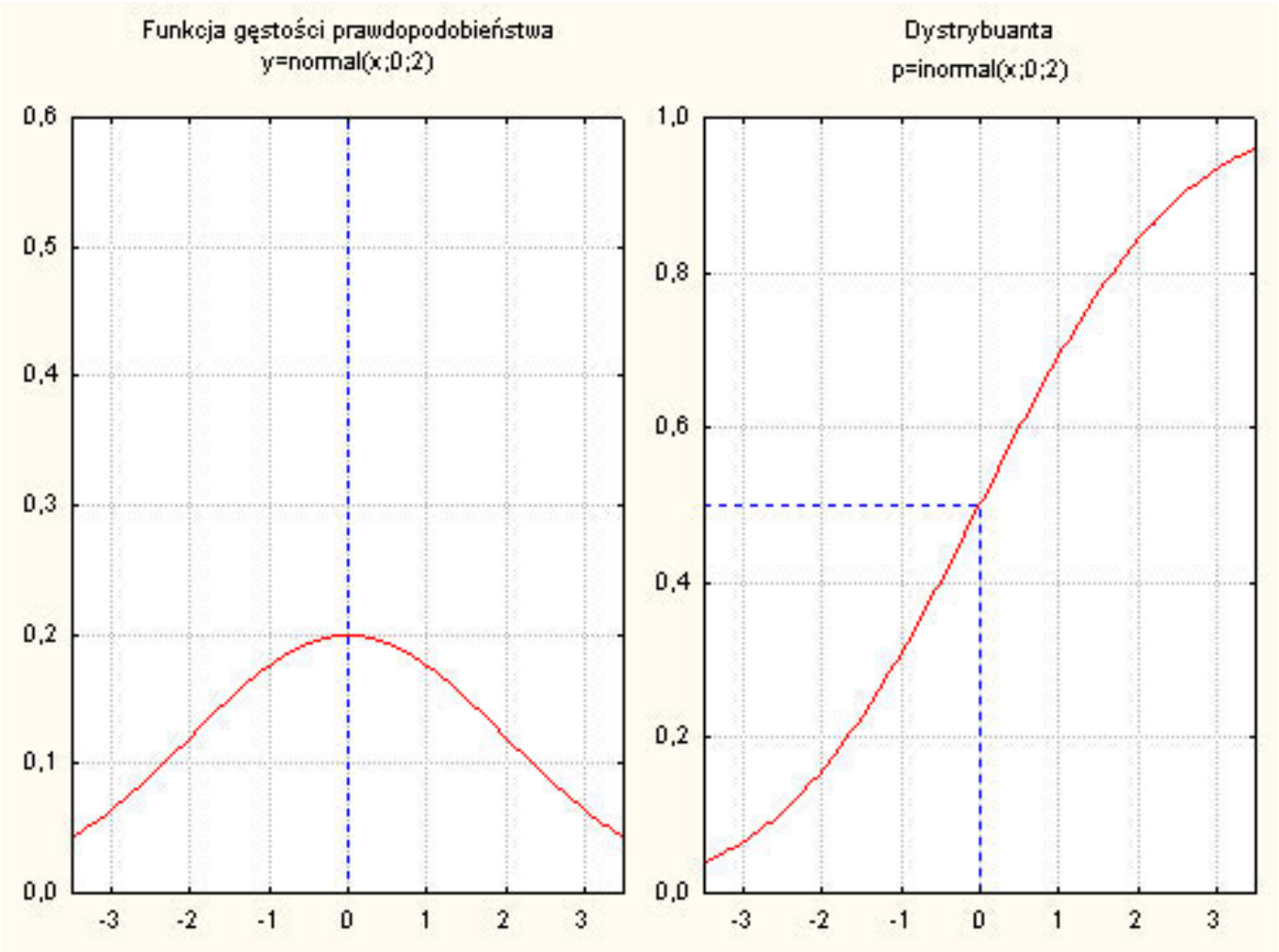
Rozkład $N(0,1)$– standardowy rozkład normalny



Rozkład $N(2,1)$


Funkcja gęstości prawdopodobieństwa
y=normal(x;2;1)
0,6
0,5
0,4
0,3
0,2
0,1
0,0
-3
-2
-1
0
1
2
3
Dystrybuanta
p=inormal(x;2;1)
1,0
0,8
0,6
0,4
0,2
0,0
-3
-2
-1
0
1
2
3

Rozkład $N(0, 2)$


Funkcja gęstości prawdopodobieństwa
y=normal(x;0;2)
0,6
0,5
0,4
0,3
0,2
0,1
0,0
-3
-2
-1
0
1
2
3
Dystrybuanta
p=inormal(x;0;2)
1,0
0,8
0,6
0,4
0,2
0,0
-3
-2
-1
0
1
2
3

Pytania

- Jaka jest różnica między cechą skokową i ciągłą — podać przykłady każdej z nich.
- Wymienić typy cech i podać po jednym przykładzie.
- Podać znane nazwy rozkładów cech i jakiego typu są to cechy.
- Podać dwa przykłady cech o rozkładzie dwumianowym.
- Podać dwa przykłady cech o rozkładzie normalnym.
- Zmienna losowa ma rozkład $N(20, 4)$. Ile wynosi $P\{X \in (16, 24)\}$?
- Omówić pojęcie populacji w badaniach statystycznych.
- Co to jest próba reprezentatywna?
- Jakie są zasady pobierania prób reprezentatywnych?
- Co to jest wnioskowanie statystyczne?
- Populacja i próba: wymienić przynajmniej dwie zasadnicze różnice.

Część II

Estymacja parametrów

Niech $X_1, X_2, \ldots, X_n$ oznacza próbę z populacji oraz θ parametr charakteryzujący tę populację. Na podstawie próby chcemy oszacować (przybliżyć) wartość parametru θ.

Estymator punktowy jest funkcją próby. Przybliża wartość parametru θ:

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$$

Estymacja punktowa parametrów cechy $X \sim N(\mu, \sigma^2)$

Estymator średniej — średnia arytmetyczna

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}$$

Estymator wariancji — wariancja próbkowa

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Estymator odchylenia standardowego

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

Estymacja punktowa parametru p cechy $X \sim D(p)$

Niech n oznacza liczbę obiektów wylosowanych z populacji, wśród których znalazło się k obiektów, które posiadają wyróżnioną właściwość. Przyjmując, że p oznacza prawdopodobieństwo wylosowania z populacji obiektu o wyróżnionej właściwości mamy:

$$\hat{p} = \frac{k}{n}$$

Uwaga. Przyjmując dla $i = 1, 2, \ldots, n$, że $P(X_i = 1) = p = 1 - P(X_i = 0)$, mamy $\hat{p} = \bar{X}$.

Przedział ufności (estymator przedziałowy) jest przedziałem o końcach zależnych od próby, który z pewnym z góry zadanym prawdopodobieństwem $1-\alpha$ pokrywa nieznaną wartość parametru θ:

$$P\{\theta \in (\underline{\theta}(X_1,\ldots,X_n), \overline{\theta}(X_1,\ldots,X_n))\} = 1-\alpha.$$

Poziom ufności jest to ustalone prawdopodobieństwo $1-\alpha$.

Populacja z wyróżnioną cechą X

Przedział ufności dla średniej μ w rozkładzie normalnym $N(\mu, \sigma^2)$

Wariancja σ^2 jest nieznana

Poziom ufności: $1 - \alpha$

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

$t(\alpha;\nu)$ jest stablicowaną wartością krytyczną rozkładu t (t–Studenta) z ν stopniami swobody.

	Dwustronne wartości krytyczne Rozkładu $t-$ Studenta			
	α			
ν	0.100	0.050	0.025	0.010
8	1.8595	2.3060	2.7515	3.3554
9	1.8331	2.2622	2.6850	3.2498
10	1.8125	2.2281	2.6338	3.1690

Przykład. Na podstawie próby 1.1, 1.2, 0.8, 0.9, 1.2, 1.3, 1.0, 0.7, 0.8, 1.0 oszacować wartość średnią μ rozkładu obserwowanej cechy $X \sim N(\mu, \sigma^2)$, na poziomie ufności $1 - \alpha = 0.95$.

$$\bar{x} = \frac{1.1 + 1.2 + \cdots + 1.0}{10} = 1.0$$

$$\sum (x_i - \bar{x})^2 = (1.1 - 1.0)^2 + \cdots + (1.0 - 1.0)^2 = 0.36$$

$$s^2 = \frac{0.36}{10 - 1} = 0.04, \; s = \sqrt{s^2} = 0.2$$

$$t(0.05; 9) = 2.2622$$

$$t(0.05; 9)\frac{s}{\sqrt{n}} = 2.2622\frac{0.2}{\sqrt{10}} = 0.14$$

$$(1 - 0.14, 1 + 0.14) = (0.86, 1.14)$$

Wniosek. Średnia wartość cechy jest jakąś liczbą z przedziału $(0.86, 1.14)$. Zaufanie do tego wniosku wynosi 95%.

Przedział ufności dla wariancji w rozkładzie normalnym

Średnia μ jest nieznana

Poziom ufności: $1 - \alpha$

$$\left(\frac{\sum_i (X_i - \bar{X})^2}{\chi^2(\frac{\alpha}{2}; n-1)}, \frac{\sum_i (X_i - \bar{X})^2}{\chi^2(1 - \frac{\alpha}{2}; n-1)} \right)$$

$\chi^2(\alpha;\nu)$ jest stablicowaną wartością krytyczną rozkładu chi–kwadrat z ν stopniami swobody.

ν	Wartości krytyczne $\chi^2(\alpha;r)$ α			
	0.975	0.950	0.050	0.025
8	2.1797	2.7326	15.5073	17.5345
9	2.7004	3.3251	16.9190	19.0228
10	3.2470	3.9403	18.3070	20.4832

Przykład. Na podstawie próby 1.1, 1.2, 0.8, 0.9, 1.2, 1.3, 1.0, 0.7, 0.8, 1.0 oszacować zróżnicowanie rozkładu obserwowanej cechy.

$$\bar{x} = \frac{1.1 + 1.2 + \cdots + 1.0}{10} = 1.0$$

$$\sum_i (x_i - \bar{x})^2 = (1.1 - 1.0)^2 + \cdots + (1.0 - 1.0)^2 = 0.36$$

$$s^2 = \frac{0.36}{10 - 1} = 0.04, \; s = \sqrt{s^2} = 0.2$$

Poziom ufności $1-\alpha = 0.95$, czyli $\alpha = 0.05$.

$$\chi^2(\frac{\alpha}{2}; n-1) = \chi^2(0.025; 9) = 19.0228$$

$$\chi^2(1-\frac{\alpha}{2}; n-1) = \chi^2(0.975; 9) = 2.7004$$

$$\left(\frac{0.36}{19.0228}, \frac{0.36}{2.7004}\right) = (0.019, 0.133)$$

Wniosek. Wariancja cechy jest liczbą z przedziału $(0.019, 0.133)$. Zaufanie do tego wniosku wynosi 95%.

Estymacja prawdopodobieństwa sukcesu

Przedział przybliżony

$$\left(\hat{p} - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

	Kwantyle u_α rozkładu normalnego $N(0,1)$				
α	0.002	0.003	0.004	0.005	0.006
0.96	1.7744	1.7866	1.7991	1.8119	1.8250
0.97	1.9110	1.9268	1.9431	1.9600	1.9774
0.98	2.0969	2.1201	2.1444	2.1701	2.1973
0.99	2.4089	2.4573	2.5121	2.5758	2.6521

Na przykład $u_{0.975} = 1.96$

Populacja 1, cecha X_1

Populacja 2, cecha X_2

Oznaczenia

Próby: $X_{11}, \ldots, X_{1n_1}$; $X_{21}, \ldots, X_{2n_2}$

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad s_i^2 = \frac{\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{n_i - 1}$$

$$s_e^2 = \frac{\sum_{j=1}^{n_1} (X_{1j} - \bar{X}_1)^2 + \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)^2}{n_1 + n_2 - 2}, \quad s_r^2 = s_e^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Ocena różnicy między średnimi $\mu_1 - \mu_2$

Ocena punktowa: $\bar{X}_1 - \bar{X}_2$

Założenia:

1. $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$
2. X_1, X_2 są niezależne
3. $\sigma_1^2 = \sigma_2^2$

Przedział ufności (poziom ufności $1 - \alpha$)

$$(\bar{X}_1 - \bar{X}_2 - t(\alpha; n_1 + n_2 - 2)s_r, \bar{X}_1 - \bar{X}_2 + t(\alpha; n_1 + n_2 - 2)s_r)$$

Przykład. Z dwóch populacji pobrano próby: 60, 62, 65, 63, 60 oraz 58, 53, 57, 56, 61. Ocenić różnicę średnich.

$$\bar{x}_1 = 62, \sum_{i=1}^{5}(x_{1i} - \bar{x}_1)^2 = 18, \bar{x}_2 = 57, \sum_{i=1}^{5}(x_{2i} - \bar{x}_2)^2 = 34$$

$$s_r^2 = \frac{18 + 34}{5 + 5 - 2}\left(\frac{1}{5} + \frac{1}{5}\right) = 2.6$$

$$t(0.05; 8) = 2.3060; t(0.05; 8)s_r = 3.72$$

$$(62 - 57 - 3.72, 62 - 57 + 3.72) = (1.28, 8.72)$$

Wniosek. Różnica średnich jest liczbą z przedziału $(1.28, 8.72)$

Ocena różnicy frakcji $p_1 - p_2$

Założenia: $X_1 \sim D(p_1), \quad X_2 \sim D(p_2)$
Cechy X_1, X_2 są niezależne

Próba 1: $X_{11}, X_{12}, \ldots, X_{1n_1}$ $(X_{1i} = 0 \text{ lub } 1)$
Próba 2: $X_{21}, X_{22}, \ldots, X_{2n_2}$ $(X_{2i} = 0 \text{ lub } 1)$

$k_1 = \sum_{i=1}^{n_1} X_{1i}$
$k_2 = \sum_{i=1}^{n_2} X_{2i}$

Ocena punktowa: $\hat{p}_1 - \hat{p}_2 = \frac{k_1}{n_1} - \frac{k_2}{n_2}$

Przybliżony przedział ufności (poziom ufności $1 - \alpha$)

$$\hat{p}_1 - \hat{p}_2 \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Iloraz frakcji: $\frac{p_1}{p_2}$ (ryzyko względne)

$$\ln(\frac{\hat{p}_1}{\hat{p}_2}) \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{1-\hat{p}_1}{n_1\hat{p}_1} + \frac{1-\hat{p}_2}{n_2\hat{p}_2}}$$

Przykład: Porównanie lekarstw ze względu na odsetek osób, które nie reagują na podany lek

p_1	p_2	$p_1 - p_2$	p_1/p_2
0.01	0.001	0.009	10
0.410	0.401	0.009	1.02

Rozkład prawdopodobieństwa oraz dane

	Y	
X	p_{11}	p_{12}
	p_{21}	p_{22}

	Y	
X	n_{11}	n_{12}
	n_{21}	n_{22}

Iloraz szans $\theta = \frac{p_{11}/p_{12}}{p_{21}/p_{22}}$

Estymator ilorazu szans $\hat{\theta} = \frac{\hat{p}_{11}/\hat{p}_{12}}{\hat{p}_{21}/\hat{p}_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$

Przedział ufności dla $\ln(\theta)$

$$\ln(\hat{\theta}) \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Pytania

- Co to jest estymator?
- Co to znaczy, że estymator jest precyzyjny?
- Podać przynajmniej dwa różne oszacowania średniej wartości cechy?
- Co to jest przedział ufności?
- Co to jest poziom ufności?
- Jaka jest interpretacja poziomu ufności?
- Od jakich czynników i jak zależy długość przedziału ufności? Czy prowadzący doświadczenie może mieć wpływ na długość przedziału ufności?
- Na podstawie badań uzyskano dla średniej następujący przedział ufności $(2, 13)$. Czy można uznać, że średnia w populacji jest równa 7 i dlaczego?

Część III

Weryfikacja hipotez statystycznych

Hipoteza statystyczna– dowolne przypuszczenie dotyczące rozkładu prawdopodobieństwa cechy (oznaczenie H_0).

Testem hipotezy statystycznej nazywamy postępowanie mające na celu odrzucenie lub nie odrzucenie hipotezy statystycznej.

Statystyką testową nazywamy funkcję próby na podstawie której wnioskuje się o odrzuceniu lub nie hipotezy statystycznej.

Błędem I rodzaju nazywamy błąd wnioskowania polegający na odrzuceniu hipotezy, gdy w rzeczywistości jest ona prawdziwa.

Błędem II rodzaju nazywamy błąd wnioskowania polegający na nieodrzuceniu hipotezy, gdy w rzeczywistości jest ona fałszywa.

	Decyzja o hipotezie	
Hipoteza	nie odrzucić	odrzucić
prawdziwa	prawidłowa	błędna
fałszywa	błędna	prawidłowa

Błąd I rodzaju kontroluje się przez zadanie małej wartości dla poziomu istotności. **Poziom istotności** jest to górne ograniczenie prawdopodobieństwa popełnienia błędu I rodzaju.

Błędu II rodzaju nie można kontrolować w taki sposób, jak błąd I rodzaju. W praktyce nie wiadomo, ile dokładnie wynosi prawdopodobieństwo popełnienia tego błędu.

Cecha X ma rozkład normalny $N(\mu, \sigma^2)$
Średnia μ oraz wariancja σ^2 są nieznane

$$H_0 : \ \mu = \mu_0$$

Test Studenta (poziom istotności α)

Próba: $X_1, \ldots, X_n$

Statystyka testowa

$$t_{\text{emp}} = \frac{\bar{X} - \mu_0}{S} \sqrt{n} \ .$$

Wartość krytyczna $t(\alpha; n-1)$

Jeżeli $|t_{\text{emp}}| > t(\alpha; n-1)$, to hipotezę odrzucamy.

Przykład. W biochemicznym doświadczeniu badano czas życia komórek w pewnym środowisku. Dokonano ośmiu pomiarów uzyskując wyniki (w godzinach): 4.7, 5.3, 4.0, 3.8, 6.2, 5.5, 4.5, 6.0. Czy można uznać, że średni czas życia komórek w badanym środowisku wynosi 4 godziny?

Cecha X — czas życia komórki ($X \sim N(\mu, \sigma^2)$)

$$H_0 : \mu = 4$$

Test Studenta; poziom istotności $\alpha = 0.05$

$$\bar{x} = 5, \ s = 0.891227, \ t_{\text{emp}} = 3.1736, \ t(0.05, 7) = 2.3646$$

Weryfikacja: Ponieważ $t_{\text{emp}} > t(0.05, 7)$, odrzucamy hipotezę

Wniosek: średni czas życia komórek w badanym środowisku nie wynosi 4 godziny.

Cecha X ma rozkład normalny $N(\mu, \sigma^2)$
Średnia μ oraz wariancja σ^2 są nieznane

$$H_0 : \sigma^2 = \sigma_0^2$$

Statystyka chi–kwadrat (poziom istotności α)
Próba: $X_1, \ldots, X_n$

Statystyka testowa $\chi^2_{\text{emp}} = \frac{\sum_i (x_i - \bar{x})^2}{\sigma_0^2}$

Wartości krytyczne $\chi^2(1 - \frac{\alpha}{2}; n-1)$, $\chi^2(\frac{\alpha}{2}; n-1)$

Jeżeli $\chi^2_{\text{emp}} < \chi^2(1 - \frac{\alpha}{2}; n-1)$ lub $\chi^2_{\text{emp}} > \chi^2(\frac{\alpha}{2}; n-1)$ to hipotezę $H_0 : \sigma^2 = \sigma_0^2$ odrzucamy.

Cecha $X \sim D(p)$
p nie jest znane

$$H_0 : \ p = p_0$$

Statystyka testowa

$$u_{\text{emp}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Wartość krytyczna: $u_{1-\frac{\alpha}{2}}$

Jeżeli $|u_{\text{emp}}| > u_{1-\frac{\alpha}{2}}$, to hipotezę odrzucamy.

Przykład. Dziesięć lat temu odsetek dzieci chorych na astmę wynosił 4%. Czy odsetek ten uległ zmianie, jeżeli w próbie dwustu dzieci rozpoznano osiemnaście przypadków astmy?

Niech X oznacza liczbę przypadków astmy wśród wylosowanych dzieci. Możemy założyć, że $X \sim B(200, p)$, gdzie p oznacza prawdopodobieństwo wylosowania dziecka chorego na astmę.
Cel: Zweryfikować hipotezę $H_0 : \ p = 0.04$

Zadaję poziom istotności $\alpha = 0.05$.

Wyznaczam $\hat{p} = 0.09$, $u_{\text{emp}} = \frac{0.09-0.04}{\sqrt{\frac{0.04(1-0.04)}{200}}} = 2.887$, $u_{0.975} = 1.96$

Ponieważ $|u_{\text{emp}}| > u_{0.975}$, hipotezę odrzucamy.

Wniosek: Odsetek dzieci chorych na astmę uległ zmianie.

Cecha X_1 ma rozkład normalny $N(\mu_1, \sigma_1^2)$
Cecha X_2 ma rozkład normalny $N(\mu_2, \sigma_2^2)$
Średnia μ_1 oraz wariancja σ_1^2 są nieznane
Średnia μ_2 oraz wariancja σ_2^2 są nieznane
$\sigma_1^2 = \sigma_2^2$

$$H_0 : \mu_1 = \mu_2$$

test t–Studenta

$$t_{\text{emp}} = \frac{\bar{X}_1 - \bar{X}_2}{S_r}$$

Wartość krytyczna $t(\alpha; n_1 + n_2 - 2)$
Jeżeli $|t_{\text{emp}}| > t(\alpha; n_1 + n_2 - 2)$, to hipotezę $H_0 : \mu_1 = \mu_2$ odrzucamy

Cecha X_1 ma rozkład dwupunktowy $D(p_1)$
Cecha X_2 ma rozkład dwupunktowy $D(p_2)$

$$H_0 : \ p_1 = p_2$$

Statystyka testowa

$$u_{\text{emp}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

gdzie

$$\hat{p}_1 = \frac{k_1}{n_1}, \ \hat{p}_2 = \frac{k_2}{n_2}, \ \hat{p} = \frac{(k_1 + k_2)}{(n_1 + n_2)}$$

Jeżeli $|u_{\text{emp}}| \geq u_{1-\alpha/2}$, to hipotezę $H_0 : p_1 = p_2$ odrzucamy

Wprowadzenie

Przykład.

- Wybrano pięciu pacjentów, którzy w równym stopniu cierpieli na pewną chorobę. Trzech z nich wybrano losowo do grupy eksperymentalnej i poddano nowej kuracji.
- Po pewnym czasie wszystkich sklasyfikowano w zależności od stopnia zaawansowania choroby.
- Pacjent, którego sklasyfikowano jako najciężej chorego, otrzymał rangę 1, drugi w kolejności rangę 2 i analogicznie pozostali.

Wprowadzenie

Przykład.

- Wybrano pięciu pacjentów, którzy w równym stopniu cierpieli na pewną chorobę. Trzech z nich wybrano losowo do grupy eksperymentalnej i poddano nowej kuracji.
- Po pewnym czasie wszystkich sklasyfikowano w zależności od stopnia zaawansowania choroby.
- Pacjent, którego sklasyfikowano jako najciężej chorego, otrzymał rangę 1, drugi w kolejności rangę 2 i analogicznie pozostali.

Wprowadzenie

Przykład.

- Wybrano pięciu pacjentów, którzy w równym stopniu cierpieli na pewną chorobę. Trzech z nich wybrano losowo do grupy eksperymentalnej i poddano nowej kuracji.
- Po pewnym czasie wszystkich sklasyfikowano w zależności od stopnia zaawansowania choroby.
- Pacjent, którego sklasyfikowano jako najciężej chorego, otrzymał rangę 1, drugi w kolejności rangę 2 i analogicznie pozostali.

Wszystkie możliwe układy rang

Eksperymentalna Kontrolna	(3,4,5) (1,2)	(2,4,5) (1,3)	(1,4,5) (2,3)	(2,3,5) (1,4)	(1,3,5) (2,4)
Eksperymentalna Kontrolna	(2,3,4) (1,5)	(1,3,4) (2,5)	(1,2,4) (3,5)	(1,2,3) (4,5)	(1,2,5) (3,4)

- Prawdopodobieństwo każdego układu, **przy założeniu, że nowa kuracja nie ma efektu,** wynosi

$$\binom{5}{2} = \frac{1}{10}$$

- Informacja, że pacjenci cierpią w równym stopniu na pewną chorobą nie wpływa na to prawdopodobieństwo.

Wszystkie możliwe układy rang

Eksperymentalna	(3,4,5)	(2,4,5)	(1,4,5)	(2,3,5)	(1,3,5)
Kontrolna	(1,2)	(1,3)	(2,3)	(1,4)	(2,4)
Eksperymentalna	(2,3,4)	(1,3,4)	(1,2,4)	(1,2,3)	(1,2,5)
Kontrolna	(1,5)	(2,5)	(3,5)	(4,5)	(3,4)

- Prawdopodobieństwo każdego układu, **przy założeniu, że nowa kuracja nie ma efektu,** wynosi

$$\binom{5}{2} = \frac{1}{10}$$

- Informacja, że pacjenci cierpią w równym stopniu na pewną chorobą nie wpływa na to prawdopodobieństwo.

Wszystkie możliwe układy rang

Eksperymentalna	(3,4,5)	(2,4,5)	(1,4,5)	(2,3,5)	(1,3,5)
Kontrolna	(1,2)	(1,3)	(2,3)	(1,4)	(2,4)
Eksperymentalna	(2,3,4)	(1,3,4)	(1,2,4)	(1,2,3)	(1,2,5)
Kontrolna	(1,5)	(2,5)	(3,5)	(4,5)	(3,4)

- Prawdopodobieństwo każdego układu, **przy założeniu, że nowa kuracja nie ma efektu,** wynosi

$$\binom{5}{2} = \frac{1}{10}$$

- Informacja, że pacjenci cierpią w równym stopniu na pewną chorobą nie wpływa na to prawdopodobieństwo.

Ogólnie. Przypuśćmy, że mamy danych N obiektów. Spośród nich wybieramy losowo n do grupy eksperymentalnej. Pozostałych $N - n$ obiektów trafia do grupy kontrolnej. Niech $S_1 < S_2 < \ldots < S_n$ będą rangami grupy eksperymentalnej. Wówczas

$$P_H(S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n) = 1/\binom{N}{n},$$

gdzie $P_H(\cdot)$ oznacza prawdopodobieństwo przy założeniu, że hipoteza H : *brak efektu grupy,* jest prawdziwa.

Test Wilcoxona

W celu zbadania wpływu czynnika (na przykład efektu nowej kuracji) rozdzielamy losowo N obiektów do dwóch grup: n do eksperymantalnej, $N-n$ do kontrolnej. Po zakończeniu badań nadajemy obiektom rangi. Niech $S_1 < S_2 < \ldots < S_n$ będą oznaczać rangi grupy eksperymentalnej. Hipotezę o braku wpływu czynnika odrzucimy na korzyść alternatywy, że jest efekt pozytywny, jeżeli rangi tej grupy okażą się istotnie duże, tzn. gdy

$$W_s := S_1 + S_2 + \ldots + S_n \geq c,$$

gdzie c jest wartością krytyczną wyznaczoną z równania

$$P_H(W_s \geq c) = \alpha$$

($\alpha-$poziom istotności)

Dla przykładu o pacjentach wyznaczymy rozkład zmiennej losowej W_s oraz wartość c dla $\alpha = 0.1$

(S_1, S_2, S_3)	(3,4,5)	(2,4,5)	(1,4,5)	(2,3,5)	(1,3,5)
W_s	12	11	10	10	9
(S_1, S_2, S_3)	(2,3,4)	(1,3,4)	(1,2,4)	(1,2,3)	(1,2,5)
W_s	9	8	7	6	8

w	6	7	8	9	10	11	12
$P_H(W_s = w)$	0.1	0.1	0.2	0.2	0.2	0.1	0.1

$$P_H(W_s \geq 12) = P_H(W_s = 12) = 0.1$$

Zatem hipotezę H odrzucamy tylko wtedy, gdy $W_s = 12$.

Statystyka Manna-Withneya

Niech $X_1, X_2, \ldots, X_m$ będzie próbą kontrolną oraz $Y_1, Y_2, \ldots, Y_n$ próbą eksperymentalną.

Niech W_{XY} – liczba par (X_i, Y_j) dla których zachodzi $X_i < Y_j$

Wówczas zachodzi

$$W_{XY} = W_s - \frac{1}{2}n(n+1)$$

Statystykę W_{XY} nazywamy statystyką Manna-Withneya

Przykład. Chcemy sprawdzić czy zniechęcanie negatywnie wpływa na wynik testu

- Spośród dziesięciu ochotników losujemy pięciu i przydzielamy do grupy kontrolnej. Pozostałych do grupy eksperymentalnej.
- Każdej osobie dajemy test A do rozwiązania i po dwóch tygodniach test B. Grupę eksperymentalną krytykujemy przy rozwiązywaniu testu B.
- Obserwujemy różnicę: wynik A-wynik B.

Przykład. Chcemy sprawdzić czy zniechęcanie negatywnie wpływa na wynik testu

- Spośród dziesięciu ochotników losujemy pięciu i przydzielamy do grupy kontrolnej. Pozostałych do grupy eksperymentalnej.
- Każdej osobie dajemy test A do rozwiązania i po dwóch tygodniach test B. Grupę eksperymentalną krytykujemy przy rozwiązywaniu testu B.
- Obserwujemy różnicę: wynik A-wynik B.

Przykład. Chcemy sprawdzić czy zniechęcanie negatywnie wpływa na wynik testu

- Spośród dziesięciu ochotników losujemy pięciu i przydzielamy do grupy kontrolnej. Pozostałych do grupy eksperymentalnej.
- Każdej osobie dajemy test A do rozwiązania i po dwóch tygodniach test B. Grupę eksperymentalną krytykujemy przy rozwiązywaniu testu B.
- Obserwujemy różnicę: wynik A-wynik B.

Przykład. Chcemy sprawdzić czy zniechęcanie negatywnie wpływa na wynik testu

- Spośród dziesięciu ochotników losujemy pięciu i przydzielamy do grupy kontrolnej. Pozostałych do grupy eksperymentalnej.
- Każdej osobie dajemy test A do rozwiązania i po dwóch tygodniach test B. Grupę eksperymentalną krytykujemy przy rozwiązywaniu testu B.
- Obserwujemy różnicę: wynik A-wynik B.

Wyniki

Eksperymentalna	5	0	16	2	9
Kontrolna	6	-5	-6	1	4

Rangi

Eksperymentalna	7	3	10	5	9
Kontrolna	8	2	1	4	6

Implementacja w pakiecie R

```
eksperymentalna<–c(5, 0, 16, 2, 9)
kontrolna<–c(6, -5, -6, 1, 4)
wilcox.test(eksperymentalna,kontrolna,alternative="greater",exact=TRUE)
```

Wydruk z R

Wilcoxon rank sum test
data: eksperymentalna and kontrolna
W = 19, p-value = 0.1111
alternative hypothesis: true location shift is greater than 0

Sprawdzić, że wartość statystyki **W** w procedurze **wilcox.test** jest wartością statystyki Manna Whithney'a

Przybliżenie rozkładem normalnym

$$E(W_s) = \frac{1}{2}n(N+1);\ D^2(W_s) = \frac{1}{12}mn(N+1)$$

z poprawką na ciągłość:

$$P(W_s \leq c) \approx \Phi\left(\frac{c - \frac{1}{2}n(N+1) + \frac{1}{2}}{\sqrt{mn(N+1)/12}}\right)$$

Powiązania

Przykład

	X_1	X_2	Y_1	X_3	Y_2	Y_3
Obserwacje	2	2	4	9	9	9
Rangi	1	2	3	4	5	6
Rangi połówkowe	1.5	1.5	3	5	5	5
	$S_1^\star$	$S_2^\star$	$R_1^\star$	$S_3^\star$	$R_2^\star$	$R_3^\star$

$$W_s^\star = S_1^\star + S_2^\star + S_3^\star$$

$(s_1^\star, s_2^\star, s_3^\star)$	(1.5,1.5,3)	(1.5,1.5,5)	(1.5,3,5)	(1.5,5,5)	(3,5,5)	(5,5,5)
$P(s_1^\star, s_2^\star, s_3^\star)$	1/20	3/20	6/20	6/20	3/20	1/20

$$P(W_s^\star \geq 13) = 3/20 + 1/20 = 4/20 = 0.2$$

$$e = 3,\ d_1 = 2,\ d_2 = 1,\ d_3 = 3$$

Przybliżenie rozkładem normalnym

$$E(W_s^\star) = \frac{1}{2}n(N+1);\ D^2(W_s^\star) = \frac{1}{12}mn(N+1) - \frac{1}{12}\frac{mn\sum_{i=1}^{e}(d_i^3 - d_i)}{N(N-1)}$$

Rozkład statystyki

$$\frac{W_s^\star - E(W_s^\star)}{D(W_s^\star)}$$

zbiega do standardowego rozkładu normalnego, gdy m, n dążą do nieskończoności oraz $\max_{i=1,\ldots,e}\left(\frac{d_i}{N}\right)$ jest oddzielone od jedynki, gdy $N \to \infty$

Alternatywa dwustronna

- Za pomocą prezentowanego testu Wilcoxona porównywaliśmy nowy zabieg ze standardowym. Porównanie to jest obciążone na korzyść zabiegu standardowego, tzn. prawdopodobieństwo decyzji na korzyść zabiegu standardowego (w sytuacji, gdy nie ma różnic między standardowym a nowym) wynosi $1 - \alpha$ (zwykle 0.9 lub więcej)
- Naszym celem jest teraz podjęcie decyzji, czy dwa zabiegi w ogóle się różnią. W tej sytuacji rozważane zabiegi odgrywają symetryczną rolę

Alternatywa dwustronna

- Za pomocą prezentowanego testu Wilcoxona porównywaliśmy nowy zabieg ze standardowym. Porównanie to jest obciążone na korzyść zabiegu standardowego, tzn. prawdopodobieństwo decyzji na korzyść zabiegu standardowego (w sytuacji, gdy nie ma różnic między standardowym a nowym) wynosi $1 - \alpha$ (zwykle 0.9 lub więcej)
- Naszym celem jest teraz podjęcie decyzji, czy dwa zabiegi w ogóle się różnią. W tej sytuacji rozważane zabiegi odgrywają symetryczną rolę

Mamy dwa zabiegi: A oraz B. $N = m + n$ obiektów dzielimy losowo na dwie grupy. Na grupie o liczności m stosujemy zabieg A, a na grupie o liczności n zabieg B. Niech W_A oraz W_B oznaczają odpowiednie sumy rang. Hipotezę H o braku różnic między zabiegami odrzucamy wtedy, gdy wartość statystyki W_B jest zbyt duża lub zbyt mała:

Dobieramy c tak, aby

$$P_H(|W_B - \frac{1}{2}n(N+1)| \geq c) = \alpha$$

lub równoważnie

$$P(|W_{XY} - \frac{1}{2}mn| \geq c) = \alpha$$

Interpretacja α: Prawdopodobieństwo błędnego uznania różnicy między zabiegami.

Wybór lepszego zabiegu

Dwa zabiegi są „nowe". Chcemy podjąć decyzję, który zabieg jest lepszy. Nie chcemy jednak faworyzować jednego z niech.

$$\begin{cases} \text{Wybieramy } B, & \text{jeżeli } W_B \geq \frac{1}{2}n(N+1)+c \\ \text{Wybieramy } A, & \text{jeżeli } W_B \leq \frac{1}{2}n(N+1)-c \\ \text{Zawieszamy decyzję}, & \text{jeżeli } |W_B - \frac{1}{2}n(N+1)| < c \end{cases}$$

Wybór c:

$$\alpha = P_H(W_B \leq \frac{1}{2}n(N+1)-c) = P_H(W_B \geq \frac{1}{2}n(N+1)+c)$$

Interpretacja α: Maksymalne prawdopodobieństwo błędnego wyboru gorszego zabiegu.

Porównywano dwie diety. W tym celu siedmiu szczurom przydzielono dietę A oraz pięciu dietę B (losowo). Obserwowano ubytek wagi. Po siedmiu tygodniach uzyskano wyniki:

A	156	183	120	113	138	145	142
B	130	148	117	133	140		

$$W_B = 2 + 4 + 5 + 7 + 10 = 28$$

Ponieważ 28 jest mniejsze od $\frac{1}{2}n(N+1) = 32.5$, dane „faworyzują" A
Poziom krytyczny: $P[W_B \leq 28] = P[W_{XY} \leq 13] = 0.265$

Model populacyjny

$N = n + m$ obiektów losujemy z populacji, n przydzielamy do grupy eksperymentalnej, m do kontrolnej. Przydział jest losowy. Wynik obiektu eksperymentalnego oznaczamy przez Y, a wynik obiektu kontrolnego przez X. Zatem X, Y są niezależnymi zmiennymi losowymi o dystrybuantach odpowiednio F oraz G:

$$F(x) = P(X \leq x), \quad G(y) = P(Y \leq y)$$

Hipotezę o tym, że zabieg eksperymentalny nie ma wpływu zapisujemy w formie:

$$H: \ F = G$$

Twierdzenie

Niech $X_1, \ldots, X_m; Y_1, \ldots, Y_n$ będą niezależnymi zmiennymi losowymi o tym samym ciągłym rozkładzie F. Niech $S_1 < \ldots < S_n$ oznaczają rangi $Y-$ów przy wspólnym rangowaniu wszystkich $N = n + m$ obiektów. Wtedy

$$P_H(S_1 = s_1, \ldots, S_n = s_n) = \frac{1}{\binom{N}{n}}$$

dla każdego $(s_1, \ldots, s_n) \in N^n : \ 1 \leq s_1 < \ldots < s_n \leq N$

Modele

- **Model losowy do porównania dwóch zabiegów.** Mamy N danych obiektów. Przydzielamy losowo m do grupy kontrolnej (grupa pod działaniem pierwszego zabiegu) oraz m do grupy eksperymentalnej (grupa pod działaniem drugiego zabiegu).
- **Model populacyjny (porównanie dwóch zabiegów)**. W sposób całkowicie losowy wybieramy z populacji N obiektów. Również losowo przydzielamy m do grupy kontrolnej oraz m do grupy eksperymentalnej.
- **Porównanie dwóch atrybutów lub dwóch podpopulacji poprzez losowanie z każdej podpopulacji.** W sposób całkowicie losowy wybieramy m obiektów z pierwszej podpopulacji (wyznaczonej przez pierwszy atrybut) oraz n obiektów z drugiej podpopulacji (wyznaczonej przez drugi atrybut)
- **Porównanie dwóch atrybutów lub dwóch podpopulacji poprzez losowanie z całej populacji.** W sposób całkowicie losowy wybieramy z populacji N obiektów. Okazuje się, że m obiektów należy do pierwszej podpopulacji, a pozostałych n do drugiej.

Postać danych (próba)

Klasa	Liczebność
1	n_1
2	n_2
$\vdots$	$\vdots$
k	n_k

Oznaczenia

$F-$ ustalony rozkład prawdopodobieństwa

$$n_i^t = Np_i^t, \ N = \sum_{i=1}^{k} n_i,$$

$$p_i^t = P_F\{X \text{ przyjęła wartość z klasy } i\}$$

Hipoteza

$$H_0 : \ X \sim F$$

Statystyka testowa

$$\chi^2_{\text{emp}} = \sum_{i=1}^{k} \frac{(n_i - n_i^t)^2}{n_i^t}$$

Wartość krytyczna $\chi^2(\alpha; k-u-1)$ (u jest liczbą nieznanych parametrów hipotetycznego rozkładu F)

Wniosek. Jeżeli $\chi^2_{\text{emp}} > \chi^2(\alpha; k-u-1)$, to hipotezę H_0 odrzucamy

Przykład. Pracodawca przypuszcza, że liczba pracowników nieobecnych w różne dni tygodnia nie jest taka sama. Chcemy to sprawdzić na podstawie danych

Dzień tygodnia	Liczba nieobecnych
Poniedziałek	200
Wtorek	160
Środa	140
Czwartek	140
Piątek	100

Cecha X — dzień nieobecności pracownika

H_0 : X ma rozkład

Pon	Wtk	Śro	Czw	Ptk
1/5	1/5	1/5	1/5	1/5

Szereg rozdzielczy ma $k = 5$ klas. Hipotetyczny rozkład jest całkowicie określony w hipotezie, czyli $u = 0$

Wartość krytyczna $\chi^2(\alpha; k - u - 1) = \chi^2(0.05; 5 - 0 - 1) = 9.4877$

Wyznaczenie wartości statystyki χ^2_{emp}

Klasa	n_i	p_i^t	n_i^t	$(n_i - n_i^t)^2/n_i^t$
Poniedziałek	200	1/5	148	$\frac{(200-148)^2}{148} = 18.270$
Wtorek	160	1/5	148	$\frac{(160-148)^2}{148} = 0.973$
Środa	140	1/5	148	$\frac{(140-148)^2}{148} = 0.432$
Czwartek	140	1/5	148	$\frac{(140-148)^2}{148} = 0.432$
Piątek	100	1/5	148	$\frac{(100-148)^2}{148} = 15.676$
				$\chi^2_{\text{emp}} = 35.676$

Ponieważ $\chi^2_{\text{emp}} > \chi^2(0.05; 4)$, więc odrzucamy hipotezę H_0

Wniosek: Odrzucamy hipotezę o równomiernym rozkładzie nieobecności w tygodniu. Zatem przypuszczenie pracodawcy można uznać za uzasadnione

Pytania

- Co to jest hipoteza statystyczna?
- Podać przykład hipotezy statystycznej oraz przykład hipotezy, która nie jest statystyczna.
- Co rozumiemy pod pojęciem testu statystycznego?
- Jaki błąd wnioskowania nazywamy błędem I rodzaju?
- Co to jest poziom istotności?
- Jaka jest interpretacja poziomu istotności?
- Co to jest błąd II rodzaju?
- Zinterpretować wniosek: odrzucono weryfikowaną hipotezę na poziomie istotności 0.05.
- Jakie założenia muszą być spełnione, by hipotezę dotyczącą różnicy między średnimi dwóch populacji można było weryfikować testem Studenta? Jak można te założenia sprawdzić?
- Jakim testem można zweryfikować hipotezę $H_0 : \mu = \mu_0$.

Część IV

Analiza korelacji, regresja

Obserwujemy dwie cechy: X oraz Y
Obiekt $\longrightarrow (X, Y)$

- Czy cechy X oraz Y są zależne?
- Opis ilościowy zależności.
- Wnioski.

Zakładamy, że łączny rozkład cech X, Y jest normalny. Założenie to oznacza, że przypuszczalna zależność między X i Y ma charakter liniowy

Test współczynnika korelacji Pearsona

$$H_0 : \ \varrho = 0 \text{ (Cechy } X, Y \text{ są niezależne)}$$

Statystyka testowa

$$R = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_j - \bar{Y})^2}}$$

Wartość krytyczna: $r(\alpha, n)$

Hipotezę odrzucamy, jeżeli $|R| > r(\alpha, n)$.

Przy założeniu, że łączny rozkład cech X, Y jest normalny, średnia wartość cechy Y zależy liniowo od wartości cechy X:

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

Funkcję $f(x) = \beta_0 + \beta_1 x$ nazywamy liniową **funkcją regresji.** Na podstawie par obserwacji (x_i, y_i), $i = 1, 2, \ldots, n$, szacujemy parametry β_0 oraz β_1.
Oszacowane parametry oznaczamy przez $\hat{\beta}_0$ oraz $\hat{\beta}_1$.

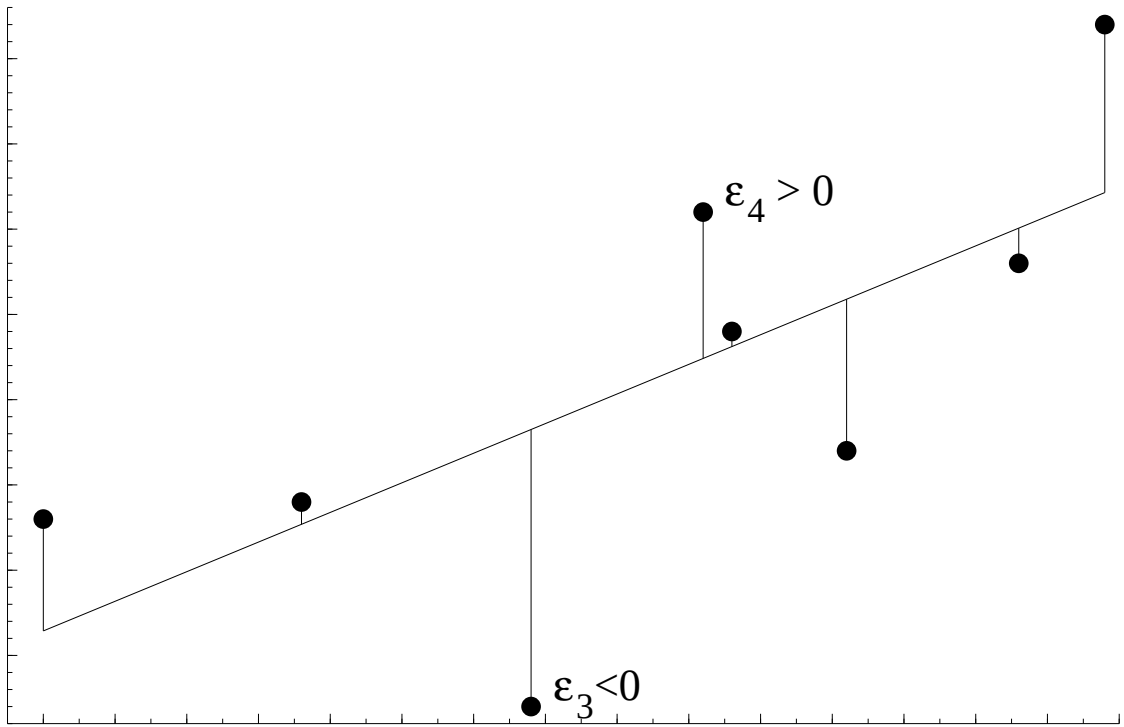
Metoda najmniejszych kwadratów

Szukamy takich parametrów β_0, β_1, które minimalizują sumę kwadratów reszt, tzn. $\sum_{i=1}^{n} \varepsilon_i^2$, gdzie reszty

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i), \quad i = 1, 1, \ldots, n$$

Tak znalezione parametry wyrażają się wzorami:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Minimalizujemy sumę kwadratów reszt: $\sum_i \varepsilon_i^2 = \text{Min!}$

Ocena przedziałowa parametrów funkcji regresji

Przedziały ufności (poziom ufności $(1-\alpha)$)

$$\beta_1 \in (\hat{\beta}_1 - t(\alpha; n-2)S_{\beta_1}, \hat{\beta}_1 + t(\alpha; n-2)S_{\beta_1})$$

$$\beta_0 \in (\hat{\beta}_0 - t(\alpha; n-2)S_{\beta_0}, \hat{\beta}_0 + t(\alpha; n-2)S_{\beta_0})$$

gdzie

$$S^2_{\beta_1} = \frac{S^2}{\sum_i (X_i - \bar{X})^2}, \quad S^2_{\beta_0} = \frac{S^2}{\sum_i (X_i - \bar{X})^2} \left(\frac{\sum_i (X_i - \bar{X})^2}{n} + \bar{X}^2 \right)$$

$$S^2 = \frac{\sum_i (Y_j - \bar{Y})^2 - \hat{\beta}_1 \sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{n-2} = \frac{\operatorname{var} Y (1 - R^2)}{n-2}$$

Przykład. Badano czy w pewnej grupie społecznej istnieje zależność między miesięcznymi wydatkami na produkty tytoniowe (X) oraz alkoholowe (Y). W tym celu wylosowano osiem rodzin uzyskując następujące wyniki:

X	69	47	73	59	76	62	87	38
Y	39	32	49	24	38	43	52	15

Populacja.

Rodziny w badanej grupie społecznej

Cechy.

X – miesięczne wydatki na produkty tytoniowe

Y – miesięczne wydatki na produkty alkoholowe

Założenie.

Cechy X oraz Y mają łączny rozkład normalny

$$H_0 : \varrho = 0$$

(nie ma zależności między X i Y)

Zadajemy poziom istotności $\alpha = 0.05$

Rachunki. $\bar{x} = 63.88, \quad \bar{y} = 36.50, \quad \sum_i (x_i - \bar{x})(y_i - \bar{y}) = 1184.50$

$$\sum_i (x_i - \bar{x})^2 = 1772.88, \quad \sum_j (y_j - \bar{y})^2 = 1086.00$$

$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_j (y_j - \bar{y})^2}} = \frac{1184.50}{\sqrt{1772.88 \cdot 1086.00}} = 0.85$$

Wartość krytyczna $r(0.05, 8) = 0.7067$

Wnioskowanie.

Ponieważ $|R| > r(0.05, 8)$, odrzucamy hipotezę zerową. Stwierdzamy, że między miesięcznymi wydatkami na produkty tytoniowe oraz alkoholowe istnieje zależność o charakterze liniowym ($\varrho \neq 0$).

Chcemy znaleźć analityczną postać tej zależności

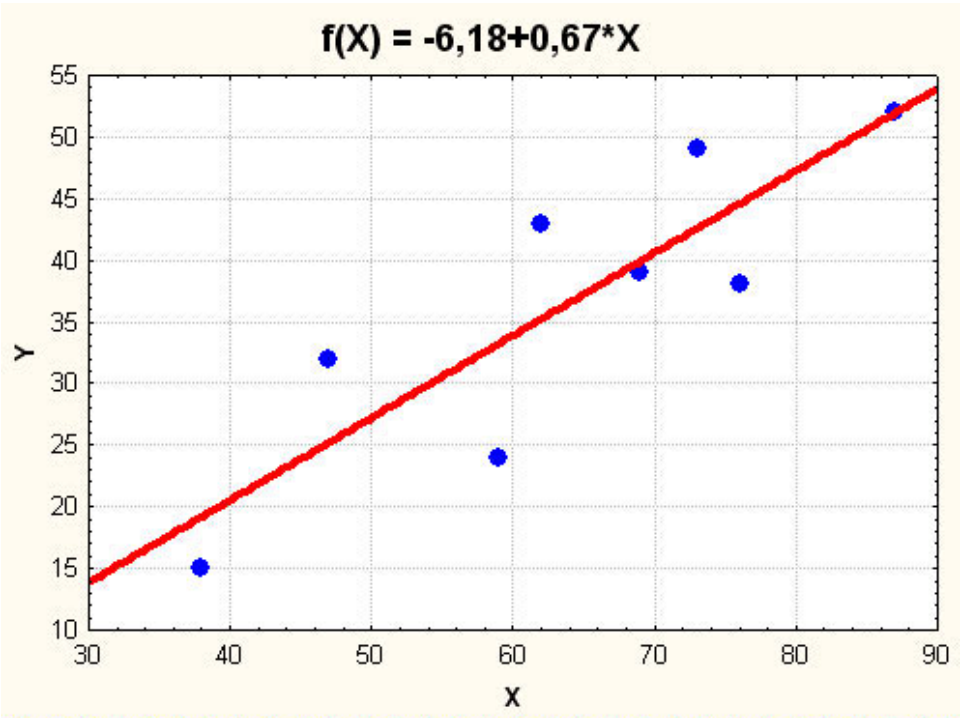
$$\hat{\beta}_1 = \frac{1184.5}{1772.88} = 0.67$$

$$\hat{\beta}_0 = 36.5 - 0.67 \cdot 63.88 = -6.18$$

$$t(0.05; 6) S_{\beta_1} = 0.41$$

Oszacowana zależność

$$\left[\begin{array}{c} \textbf{przeciętne miesięczne} \\ \textbf{wydatki na produkty} \\ \textbf{alkoholowe} \end{array}\right] = 0.67 \left[\begin{array}{c} \textbf{miesięczne} \\ \textbf{wydatki na} \\ \textbf{prod. tytoniowe} \end{array}\right] - 6.18$$



Rysunek: Regresja Y względem X

Interpretacja współczynnika kierunkowego oszacowanej funkcji regresji:

Jeżeli jakieś rodziny wydają, w stosunku do innych rodzin, o złotówkę więcej na produkty tytoniowe, to tym samym na alkohol wydają średnio o około 67 groszy więcej; błąd statystyczny tej oceny wynosi 41 groszy:

$$\beta_1 \in (0.67 - 0.41; 0.67 + 0.41).$$

Y – zmienna zależna (objaśniana)

X – zmienna niezależna (objaśniająca)

$(Y_1, x_1), (Y_2, x_2), \cdots, (Y_n, x_n)$ – próba

Model.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ \ i = 1, \ldots, n$$

Zakładamy, że $\epsilon_i, \ \ i = 1, \ldots, n$, są niezależnymi zmienymi losowymi o tym samym rozkładzie normalnym $N(0, \sigma^2)$

Estymacja punktowa.

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_i (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Estymacja przedziałowa.

$$\beta_1 \in (\hat{\beta}_1 - t(\alpha; n-2) S_{\beta_1}, \hat{\beta}_1 + t(\alpha; n-2) S_{\beta_1})$$
$$\beta_0 \in (\hat{\beta}_0 - t(\alpha; n-2) S_{\beta_0}, \hat{\beta}_0 + t(\alpha; n-2) S_{\beta_0})$$

Interpretacja oceny parametru β_1.

Jeżeli wartość x zmiennej niezależnej zwiększymy o jednostkę, to średnia wartość cechy Y zmieni się o około $\hat{\beta}_1$ jednostek, a dokładniej zmieni się o wartość z przedziału $\hat{\beta}_1 \pm t(\alpha; n-2) S_{\beta_1}$.

$$H_0 : \beta_1 = 0$$

(Y nie zależy od X)

Statystyka testowa

$$F_{\text{emp}} = \frac{\hat{\beta}_1^2}{S_{\beta_1}^2} = \frac{\hat{\beta}_1 \sum_i (x_i - \bar{x})(Y_i - \bar{Y})}{S^2}$$

Hipotezę odrzucamy, jeżeli $F_{\text{emp}} > F(\alpha; 1, n-2)$.

$F(\alpha; 1, n-2)$ jest wartością krytyczną rozkładu F.

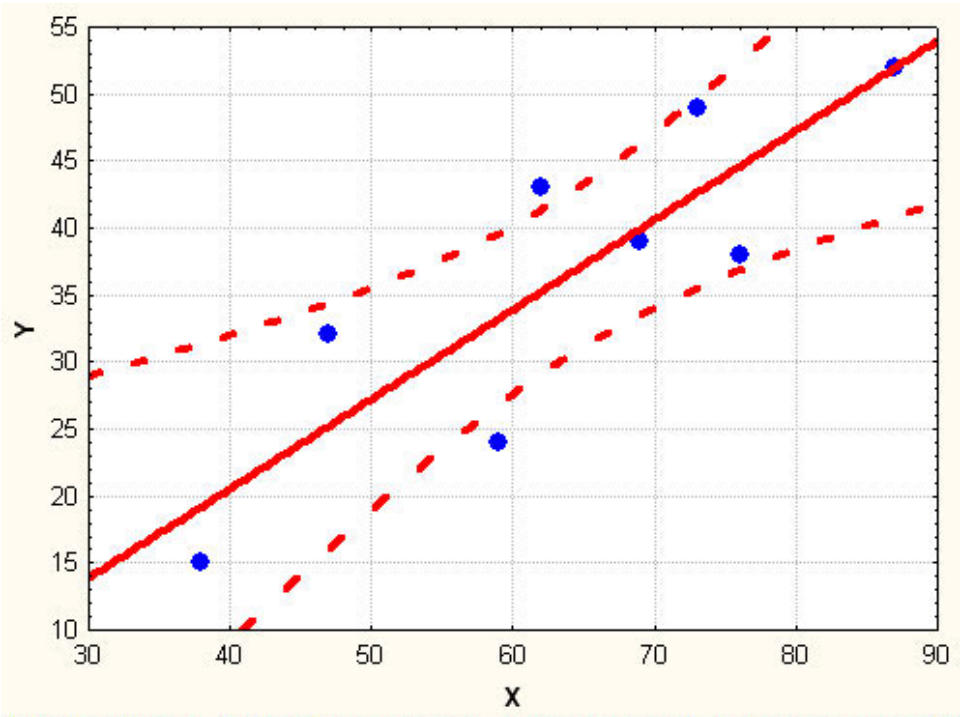
Obszar ufności dla prostej regresji umożliwia nam wnioskowanie o wartościach średnich zmiennej Y jednocześnie dla wielu wybranych wartości zmiennej X.

$$f(x) \in (\hat{f}(x) - t(\alpha; n-2)S_Y; \hat{f}(x) + t(\alpha; n-2)S_Y)$$

gdzie

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$S_Y^2 = S^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right)$$



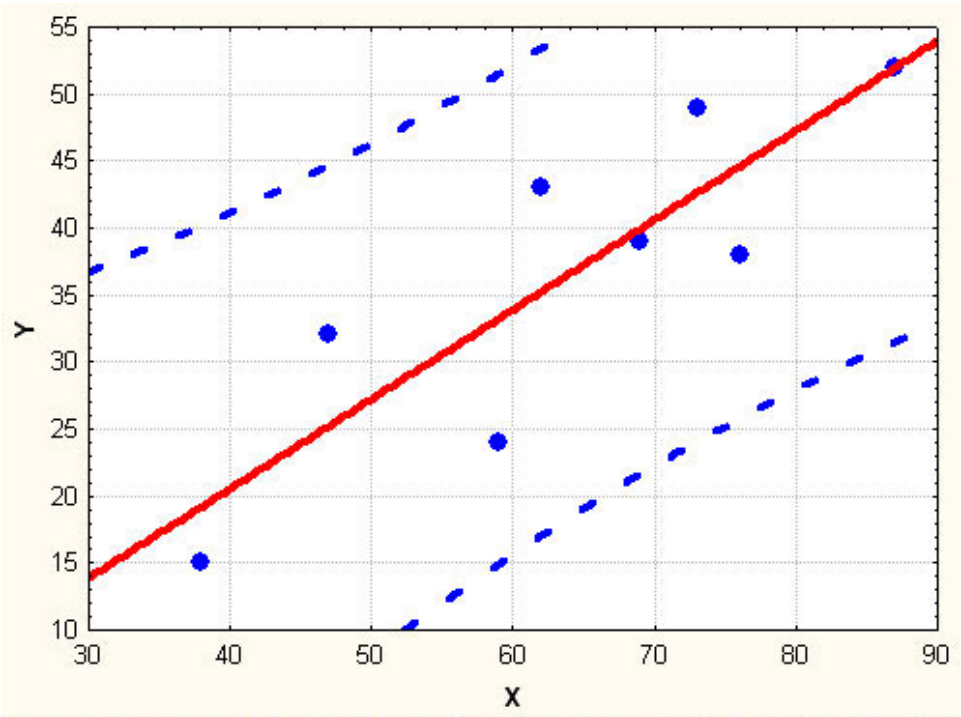
Rysunek: Obszar ufności

Obszar predykcji umożliwia nam wnioskowanie o wartościach zmiennej Y jednocześnie dla wielu wybranych wartości zmiennej X.

$$Y(x) \in (\hat{f}(x) - t(\alpha; n-2)S_{Y(x)}; \hat{f}(x) + t(\alpha; n-2)S_{Y(x)})$$

gdzie $Y(x)$ oznacza wartość zmiennej Y dla wybranej wartości x zmiennej X oraz

$$S^2_{Y(x)} = S^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)$$



Rysunek: Obszar predykcji

Część V

Testy nieparametryczne na niezależność

Obserwujemy dwie cechy: X oraz Y.

Obiekt $\longrightarrow (X, Y)$

Cechy X oraz Y są dowolnego typu.
Wartości obu cech są podzielone na klasy.

- Test nie służy do badania kierunku powiązania cech
- Zaleca się, aby próba była na tyle duża, aby liczebności teoretyczne poszczególnych klas były równe co najmniej 5

Klasy cechy Y	Klasy cechy X			
	1	2	...	m
1	n_{11}	n_{12}	...	n_{1m}
2	n_{21}	n_{22}	...	n_{2m}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
k	n_{k1}	n_{k2}	...	n_{km}

$$n_{ij}^t = \frac{n_{i\cdot} n_{\cdot j}}{N}, \quad N = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij},$$

$$n_{i\cdot} = \sum_{j=1}^{m} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$$

Hipoteza

$$H_0 : \text{Cechy } X,\ Y \text{ są niezależne}$$

α – poziom istotności
Statystyka testowa

$$\chi^2_{\text{emp}} = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - n^t_{ij})^2}{n^t_{ij}}$$

Jeżeli $\chi^2_{\text{emp}} > \chi^2(\alpha; (k-1)(m-1))$, to hipotezę H_0 odrzucamy

Przykład. Badano związek pomiędzy wykształceniem (X) a zarobkami (Y).

H_0: cechy X oraz Y są niezależne

Test chi–kwadrat niezależności ($\alpha = 0.05$)

Zbadano łącznie $N = 950$ osób

Tabela obserwacji n_{ij}

	podstawowe	średnie	wyższe	ponad wyższe
≤ 500	21	41	93	47
500-1000	33	37	35	53
1000-1500	45	75	27	43
1500-2000	30	48	50	55
≥ 2000	71	47	49	50

Liczebności brzegowe:

$$n_{1\cdot} = 21 + 41 + 93 + 47 = 202$$

$$n_{2\cdot} = 158, \;\; n_{3\cdot} = 190, \;\; n_{4\cdot} = 183, \;\; n_{5\cdot} = 217$$

$$n_{\cdot 1} = 21 + 33 + 45 + 30 + 71 = 200$$

$$n_{\cdot 2} = 248, \;\; n_{\cdot 3} = 254, \;\; n_{\cdot 4} = 248.$$

Liczebności teoretyczne:

$$n^t_{11} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N} = \frac{202 \cdot 200}{950} = 42.5263$$

$$n^t_{43} = \frac{n_{4\cdot} \cdot n_{\cdot 3}}{N} = \frac{183 \cdot 254}{950} = 48.9284$$

Tabela liczbności teoretycznych n_{ij}^t

	podstawowe	średnie	wyższe	ponad wyższe
≤ 500	42.5263	52.7326	54.0084	52.7326
500-1000	33.2632	41.2463	42.2442	41.2463
1000-1500	40.0000	49.6000	50.8000	49.6000
1500-2000	38.5263	47.7726	48.9284	47.7726
≥ 2000	45.6842	56.6484	58.0189	56.6484

Wyznaczenie $(n_{ij} - n_{ij}^t)^2/n_{ij}^t$ dla wszystkich dwudziestu kombinacji i, j.

$$\frac{(n_{11} - n_{11}^t)^2}{n_{11}^t} = \frac{(21 - 42.5263)^2}{42.5263} = 10.8964$$

Tabela wartości $\frac{(n_{ij}-n_{ij}^t)^2}{n_{ij}^t}$

	podstawowe	średnie	wyższe	ponad wyższe
≤ 500	10.8964	2.6104	28.1501	0.6232
500-1000	0.0021	0.4372	1.2423	3.3494
1000-1500	0.6250	13.0073	11.1504	0.8782
1500-2000	1.8870	0.0011	0.0235	1.0934
≥ 2000	14.0287	1.6433	1.4020	0.7803

Wartość statystyki testowej jest sumą wartości podanych w powyższej tabeli

$$\chi^2_{\text{emp}} = 93.8311$$

W tablicach znajdujemy wartość krytyczną zmiennej losowej o rozkładzie chi–kwadrat na poziomie istotności $\alpha = 0.05$ oraz dla $(4-1)(5-1) = 12$ stopni swobody:

$$\chi^2(0.05; 12) = 21.0261$$

Ponieważ $\chi^2_{\text{emp}} > \chi^2(0.05; 12)$, więc stwierdzamy, że istnieje zależność między wykształceniem i zarobkami.

Część VI

Elementy statystyki opisowej

OZNACZENIA

	Ilość		Cena jednostkowa	
artykułu	Rok 0	Rok 1	Rok 0	Rok 1
1	q_{10}	q_{11}	p_{10}	p_{11}
2	q_{20}	q_{21}	p_{20}	p_{21}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	q_{k0}	q_{1k}	p_{k0}	p_{1k}

OZNACZENIA

Numer	Wartość	Wartość	Wartość	Wartość
1	$q_{10}p_{10}$	$q_{11}p_{11}$	$q_{10}p_{11}$	$q_{11}p_{10}$
2	$q_{20}p_{20}$	$q_{21}p_{21}$	$q_{20}p_{21}$	$q_{21}p_{20}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$q_{k0}p_{k0}$	$q_{k1}p_{k1}$	$q_{k0}p_{k1}$	$q_{k1}p_{k0}$
Suma	w_{00}	w_{11}	w_{01}	w_{10}

Indeksy agregatowe

Indeks zmian wartości	$I_w = w_{11}/w_{00}$
Indeks Laspayresa zmian ilości	$I_q^L = w_{10}/w_{00}$
Indeks Laspayresa zmian cen	$I_p^L = w_{01}/w_{00}$
Indeks Paaschego zmian ilości	$I_q^P = w_{11}/w_{01}$
Indeks Paaschego zmian cen	$I_p^P = w_{11}/w_{10}$
Indeks Fishera zmian ilości	$I_q^F = \sqrt{I_q^L \cdot I_q^P}$
Indeks Fishera zmian cen	$I_p^F = \sqrt{I_p^L \cdot I_p^P}$

SZEREG ROZDZIELCZY: opisuje rozkład wartości badanej cechy (np. cechy X)

Przedział klasowy	Liczebność	Liczebność skumulowana
$x_0 - x_1$	n_1	$n_{(1)} = n_1$
$x_1 - x_2$	n_2	$n_{(2)} = n_1 + n_2$
$\vdots$	$\vdots$	$\vdots$
$x_{k-1} - x_k$	n_k	$n_{(k)} = n_1 + n_2 + \cdots + n_k (= n)$

Dla liczby p takiej, $0 \leq p \leq 1$ niech x_p, n_p, h_p oznaczają początek, liczebność i długość przedziału zawierającego obserwację o numerze $p \cdot n$ oraz niech $n_{(p)}$ oznacza liczebność skumulowaną przedziału poprzedzającego przedział o początku x_p. Niech $\dot{x}_i$ oznacza środek $i-$tego przedziału.

Mierniki położenia opisują poziom badanej cechy, tzn. w sposób syntetyczny charakteryzują wartości przyjmowane przez badaną cechę.

Mierniki położenia	
średnia $\bar{x}$	$\frac{1}{n}\sum_{i=1}^{k} \dot{x}_i n_i$
mediana *Me*	$x_{0.5} + \frac{h_{0.5}}{n_{0.5}}\left(\frac{n}{2} - n_{(0.5)}\right)$
dolny kwartyl Q_1	$x_{0.25} + \frac{h_{0.25}}{n_{0.25}}\left(\frac{n}{4} - n_{(0.25)}\right)$
górny kwartyl Q_3	$x_{0.75} + \frac{h_{0.75}}{n_{0.75}}\left(\frac{3n}{4} - n_{(0.75)}\right)$
dominanta (moda) *D*	$x_D + h_D \frac{n_D - n_{D-1}}{2n_D - n_{D+1} - n_{D-1}}$
minimum *Min*	x_0
maksimum *Max*	x_k

Mierniki rozproszenia opisują zróżnicowanie cechy, tzn. w sposób syntetyczny opisują zróżnicowanie wartości przyjmowanych przez badaną cechę

Mierniki rozproszenia	
rozstęp R	$Max - Min$
wariancja S^2	$\frac{1}{n-1}\sum_{i=1}^{k} n_i(\dot{x}_i - \bar{x})^2$
odchylenie standardowe $S = \sqrt{S^2}$	$\sqrt{\frac{1}{n-1}\sum_{i=1}^{k} n_i(\dot{x}_i - \bar{x})^2}$
odchylenie przeciętne d	$\frac{1}{n}\sum_{i=1}^{k} n_i\lvert\dot{x}_i - \bar{x}\rvert$
odchylenie ćwiartkowe Q	$\frac{Q_3 - Q_1}{2}$
współczynnik zmienności V	$\frac{S}{\bar{x}}100\%$

Przykład

Wysokość pożyczki		Liczba dłużników
0	20	155
20	50	260
50	100	215
100	200	315
200	300	55
Suma		1000

i	x_{i-1}	x_i	$\dot{x}_i$	n_i	$n_{(i)}$
1	0	20	10	155	155
2	20	50	35	260	415
3	50	100	75	215	630
4	100	200	150	315	945
5	200	300	250	55	1000
SUMA				1000	

Mierniki położenia	
średnia $\bar{x}$	$\frac{87775}{1000} = 87.775$
mediana Me	$50 + \frac{50}{215} \cdot \left(\frac{1000}{2} - 415\right) = 69.77$
dolny kwartyl Q_1	$20 + \frac{30}{260} \cdot \left(\frac{1000}{4} - 155\right) = 30.96$
górny kwartyl Q_3	$100 + \frac{100}{315} \cdot \left(\frac{3 \cdot 1000}{4} - 630\right) = 138.10$
dominanta (moda) D	$100 + 100 \cdot \frac{315-215}{2 \cdot 315-55-215} = 127.78$
minimum Min	0
maksimum Max	300

Mierniki rozproszenia

$$R = 300 - 0 = 300$$

$$S^2 = \frac{1}{999}\left(155(10 - 87.775)^2 + 260(35 - 87.775)^2 + \ldots \right.$$
$$\left. \ldots + 55(250 - 87.775)^2\right) = 7567.11$$

$$S = \sqrt{7567.11} = 86.989$$

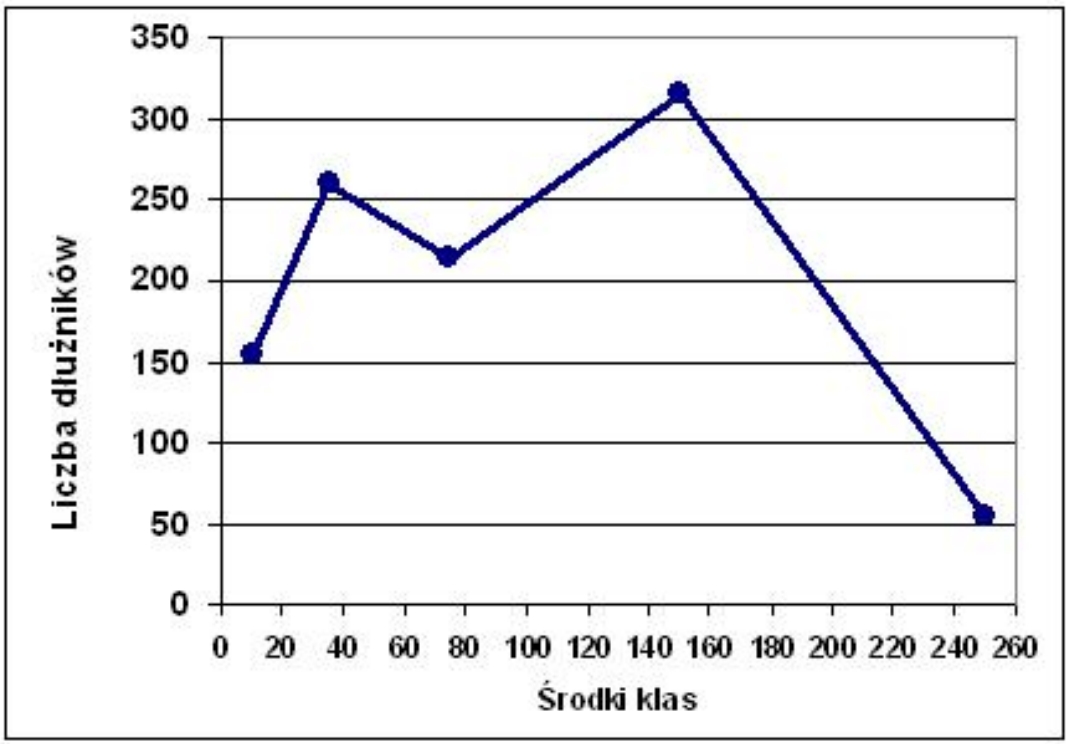
$$V = \frac{86.989}{87.775}100\% = 99.1\%$$

Jeżeli rozkład wartości cechy X jest podobny do rozkładu normalnego, to w przybliżeniu 70% wartości tej cechy

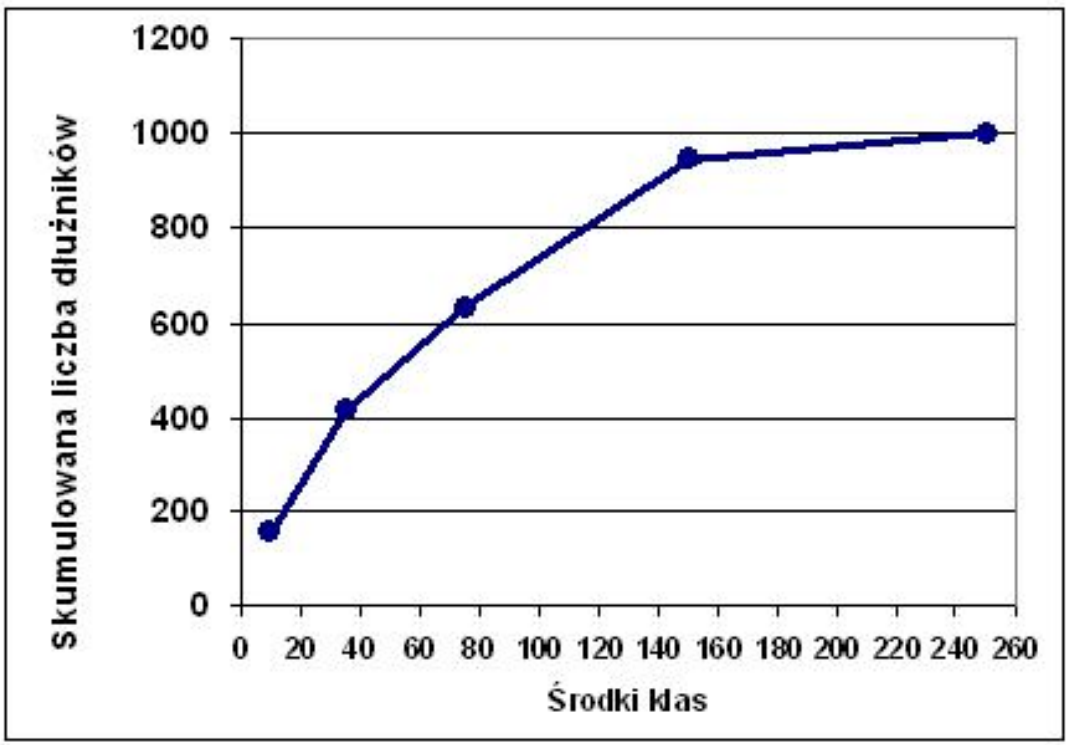
- zawiera się w przedziale
$$(\bar{x} - S, \bar{x} + S)$$
- spełnia nierówność (o ile $\bar{x} > 0$)
$$\left| \frac{x - \bar{x}}{\bar{x}} \right| 100\% < V$$

gdzie x – wartość cechy X.

Wielobok liczebności: $\{(\dot{x}_i, n_i) : \quad i = 1, 2, \ldots k\}$



Wielobok skumulowanej liczebności: $\{(\dot{x}_i, n_{(i)}) : \quad i = 1, 2, \ldots k\}$

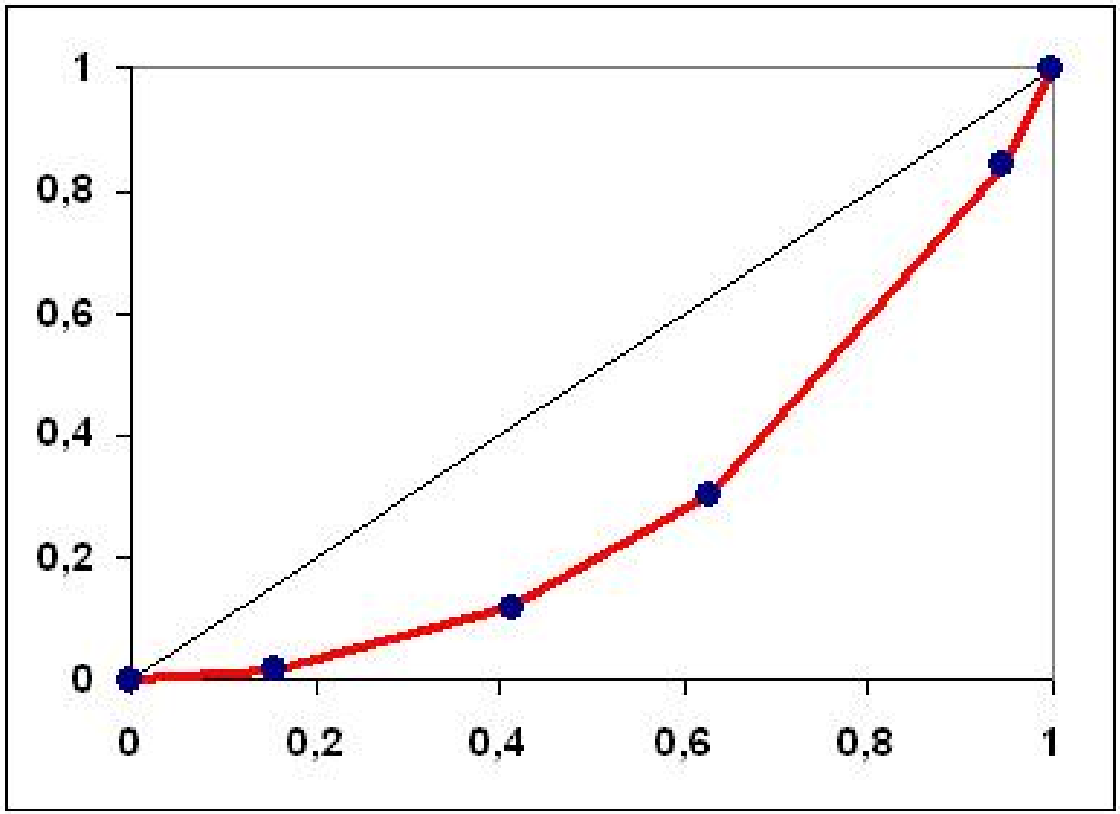


Dodatkowe wyliczenia

i	$\dot{x}_i$	n_i	$n_{(i)}$	$n_{(i)}/n$	w_i	$w_{(i)}$	$w_{(i)}/w$
1	10	155	155	0.155	1550	1550	0.018
2	35	260	415	0.415	9100	10650	0.121
3	75	215	630	0.630	16125	26775	0.305
4	150	315	945	0.945	47250	74025	0.843
5	250	55	1000	1.000	13750	87775	1.000

gdzie $w_i = n_i \dot{x}_i$, $\quad w_{(i)} = w_1 + w_2 + \ldots + w_i$ oraz $w = \sum_i w_i$

Krzywa koncentracji (Lorenza): $\{(\frac{n_{(i)}}{n}, \frac{w_{(i)}}{w}) : \quad i = 0, \ldots, k\}$



Krzywa koncentracji obrazuje nierównomierność rozdziału ogólnej sumy wartości cechy pomiędzy poszczególne jednostki zbiorowości

Oś X: skumulowany odsetek dłużników

Oś Y: skumulowany odsetek sumy pożyczek

Brak koncentracji: $\frac{n_{(i)}}{n} = \frac{w_{(i)}}{w}$, dla $i = 1, \ldots, k$.

Współczynnik koncentracji Giniego $G = \frac{a}{a+b}$